Vector Quantized Cross-lingual Unsupervised Domain Adaptation for Speech Emotion Recognition

Pravin Mote[1,2], *Donita Robinson*[3], *Elizabeth Richerson*[3], *Carlos Busso*[1]

[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Electrical and Computer Engineering, The University of Texas at Dallas, USA
[3]Laboratory for Analytic Sciences, North Carolina State University, USA

pmote@andrew.cmu.edu, {drobins7,ericher}@ncsu.edu, busso@cmu.edu

Abstract

Building *speech emotion recognition* (SER) models for low-resource languages is challenging due to the scarcity of labeled speech data. This limitation mandates the development of cross-lingual unsupervised domain adaptation techniques to effectively utilize labeled data from resource-rich languages. Inspired by the TransVQA framework, we propose a method that leverages a shared quantized feature space to enable knowledge transfer between labeled and unlabeled data across languages. The approach utilizes a quantized codebook to capture shared features, while reducing the domain gap, and aligning class distributions, thereby improving classification accuracy. Additionally, an information loss (InfoLoss) mechanism mitigates critical information loss during quantization. InfoLoss achieves this goal by minimizing the loss within the simplex of posterior class label distributions. The proposed method demonstrates superior performance compared to state-of-the-art baseline approaches.
Index Terms: Speech Emotion Recognition, Cross-lingual Unsupervised Domain Adaptation, Discrete Features, InfoLoss

1. Introduction

Emotions are fundamental to human communication [1], conveyed primarily through speech [2,3] and facial expressions [4]. However, *human-computer interaction* (HCI) [5] often lacks emotional awareness, limiting its effectiveness in applications such as personalized learning, virtual assistance, and therapy. *Speech emotion recognition* (SER) offers a promising solution to enhance HCI, but linguistic barriers hinder its applicability in multilingual settings [6–8]. The scarcity of labeled data for most languages, due to the high cost of emotional annotation, further complicates SER development, with only a few languages having sufficient resources [9–12]. We need strategies to transfer knowledge to deal with low-resource languages.

Monolingual SER models struggle with linguistically distinct languages due to differences in phonetic, syntactic, prosodic, lexical, and semantic structures [13]. Given the performance gap between closely and distantly related languages, relying on separate models becomes impractical. To address this problem, we aim to develop a unified multilingual SER system capable of handling multiple languages within a single framework. One approach is to combine datasets from various languages for supervised training, enabling the model to learn linguistic nuances [14–16]. However, due to the scarcity of labeled data, we prioritize unsupervised domain adaptation to learn linguistic variations and transfer emotional knowledge to low-resource languages.

Inspired by the TransVQA approach used for image classification [17], this study proposes a shared codebook-based unsupervised domain adaptation technique that learns a quantized shared space from the feature space of both labeled and unlabeled data, as well as from the posterior class-label distribution. The latent codebook effectively captures language-specific variations by encoding both the Euclidean feature space and the posterior label simplex into an aligned discrete representation, facilitating cross-lingual knowledge transfer. The proposed method of quantized codebook for domain adaptation operates with three distinct objectives. The first objective focuses on aligning unlabeled data with labeled data at the global domain level by minimizing the cross-domain discrepancy. The second objective focuses on class-wise alignment across both domains, where labeled data annotations and pseudo-labels for unlabeled data are utilized to reduce intra-class disparity while maintaining a higher inter-class separation. This structured alignment enhances classification performance in the unlabeled domain. While the first two objectives focus on feature alignment in the Euclidean space, the third objective operates in the posterior class-label space. It aims to mitigate information loss introduced by the quantization process by minimizing the discrepancy between the true and predicted label distributions [18].

To evaluate the effectiveness of our unsupervised domain adaptation strategy, we conduct experiments on two linguistically distinct datasets: the MSP-Podcast corpus [11] for English as the labeled dataset and the BIIC-Podcast corpus [10] for Taiwanese-Mandarin as the unlabeled dataset. These languages exhibit significant differences in phonetics, syntax, and tonal characteristics, further impacting SER performance. Experimental results demonstrate that the proposed method improves SER performance on unlabeled datasets by 14.8% compared to cross-corpus SER without adaptation. It outperforms state-of-the-art unsupervised domain adaptation methods, including ladder network and domain adversarial network, highlighting its effectiveness in bridging linguistic and domain discrepancies. Notably, the addition of InfoLoss contributes to a 3% performance gain by mitigating information loss during quantization. These findings underscore the potential of our approach for enhancing multilingual SER, paving the way for more robust and generalizable SER models.

2. Related Work

2.1. Domain Adaptation

Various domain adaptation methods have been explored, including supervised, semi-supervised, and unsupervised approaches. In supervised learning, Shami and Verhelst [14] enhanced dataset diversity by merging multiple corpora, while Hassan et al. [16] improved cross-domain generalization by assigning higher weights to critical training samples. Schuller et al. [15] similarly optimized domain adaptation by strategically selecting essential samples. In semi-supervised learning, Abdelwa-

hab and Busso [19] proposed selecting a subset of samples using active learning strategies. For unsupervised approaches, Parthasarathy and Busso [20] employed an encoder-decoder framework to learn shared representations, enhancing SER performance. Abdelwahab et al. [21] leveraged the gradient reversal method in a domain classifier to extract non-discriminatory features, facilitating adaptation. Liu and Tuzel [22] used two *generative adversarial networks* (GANs), each dedicated to a domain but sharing parameters to extract common features across domains.

Cross-lingual domain adaptation is particularly challenging due to linguistic differences. Feraru et al. [13] showed that adaptation performance declines for languages from different families. To address this problem, Upadhyay et al. [7, 8] used the most similar phonemes between two languages as anchors for knowledge transfer, improving adaptation. Amiriparian et al. [23] compiled a multi-lingual corpus from 37 existing datasets and designed an enhanced version of HuBERT to achieve better cross-lingual SER performance. Zehra et al. [24] employed ensemble learning, aggregating predictions from multiple classifiers via majority voting. Additionally, Lee [25] enhanced cross-lingual SER performance through normalization techniques and multitask learning, incorporating auxiliary tasks such as gender and language classifications.

2.2. Discrete Speech Units and InfoLoss

Discrete representations are latent codes derived by mapping features to their nearest code in a predefined set, known as a codebook, through a process called nearest-neighbor quantization [26]. Van Der Oord et al. [27] introduced the *vector quantized variational autoencoder* (VQ-VAE) framework, demonstrating the effectiveness of discrete encoded latent space. Discrete features, as opposed to continuous representations, significantly reduce computational complexity, enabling scalable models for more complex tasks. The efficacy of compact and meaningful discrete codebook has been validated in models such as VQ-GAN [28] and VQ-Diffusion [29]. Sun et al. [17] proposed the TransVQA framework, leveraging a quantized codebook for unsupervised domain adaptation in image classification. This approach incorporates global and local losses for domain and class alignment, respectively. This study serves as a building block for our formulation for cross-lingual SER.

The quantization process often results in information loss. To address this limitation, Lazebnik and Raginsky [18] proposed an information loss minimization approach that jointly optimizes feature and class probability quantization, preserving discriminative information. Our proposed method integrates a shared codebook learned from both the Euclidean feature space and the posterior class-label space, aiming to minimize cross-lingual domain shifts and enhance performance on unlabeled data. This paper adapts the TransVQA [17] approach to SER and enhance its framework by minimizing quantization error using the InfoLoss formulation [18].

3. Methodology

The features of TransVQA [17], such as the latent codebook of discrete features and loss functions dedicated to aligning domain and class distributions, are effective for unsupervised domain adaptation. The TransVQA framework's ability to learn discriminative features across multiple domains makes it well-suited for addressing cross-lingual differences in SER. Therefore, we build on this framework by adapting its alignment objectives to address the intrinsic challenges in SER. A key distinction from the TransVQA framework and our proposed approach is the inclusion of an additional objective to compensate for information loss caused by feature quantization.

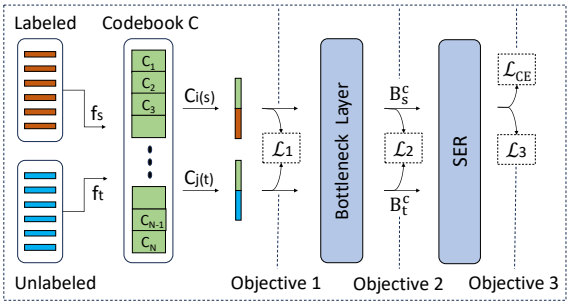

Figure 1: *Unsupervised domain adaptation using discrete speech representations. The approach has three objectives to (1) reduce domain mismatch, (2) align emotional classes, and (3) mitigate information loss due to quantization.*

Figure 1 illustrates the proposed unsupervised domain adaptation strategy that utilizes a latent codebook of discrete features to represent both labeled and unlabeled domains. The approach incorporates domain and class distribution alignment objectives, inspired by the TransVQA framework. To mitigate the quantization loss, we propose to utilize the InfoLoss [18], which aims to minimize the discrepancy between the true posterior class distribution and the predicted label distribution derived from the codebook features. With the help of InfoLoss, the latent codebook learns a shared quantized space across both domains by leveraging representations from the Euclidean feature space and the simplex of posterior class distributions in a three-objective process. The first two objectives focus on domain-level and class-level alignment, while the third objective refines the quantized space through the InfoLoss and cross-entropy loss for SER. The shared codebook captures discriminative features common to both domains, enhancing domain adaptation. It is expected to encode essential prosodic, spectral, and temporal characteristics while filtering out domain-specific variations such as speaker identity, accent, and background noise. This selective representation ensures robustness to cross-lingual variations, effectively reducing domain shifts.

We conducted the adaptation process using the wavLM [30] speech representation. The wavLM feature extraction model, obtained from the Hugging Face library, was fine-tuned on the MSP-Podcast corpus [11] with SER as the downstream task. To expedite the convergence of the codebook, we initialized it with centroids derived from the clustered feature space of both labeled and unlabeled domains. These centroids were obtained by applying k-means clustering to the wavLM feature space, incorporating frames from both domains. This initialization provides a structured starting point for the codebook, enabling it to capture shared representations more effectively.

In Figure 1, each frame of both labeled and unlabeled utterances is mapped into a corresponding vector from the codebook, which is then updated during the backpropagation step. Among the N available codes in the codebook, the nearest code for each frame is selected using the *k-nearest neighbors* (kNN) algorithm [31] and utilized in subsequent alignment objectives as described below.

3.1. Domain-Level Alignment

As described in Figures 1 and 2, the first objective of our approach aims to reduce the discrepancy between the labeled and

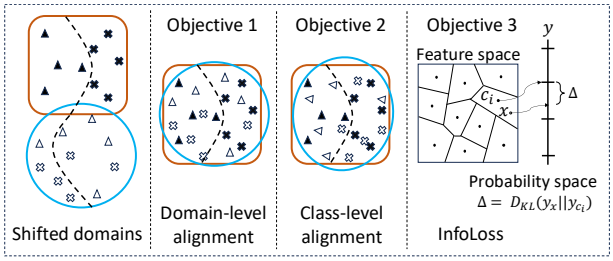

Figure 2: *Objectives for our unsupervised domain adaptation using discrete speech representations. Objective-1 reduces domain shift, Objective-2 aligns classes from different domains, Objective-3 helps to reduce quantization loss.*

unlabeled domains by minimizing the Euclidean distance between their wavLM features – the labeled domain (f_s) and the unlabeled domain (f_t) – along with their corresponding quantized codes $C_i(s)$ and $C_j(t)$, as formulated in Equation 1. The first term in Equation 1 facilitates adaptation in the unlabeled domain by leveraging prior knowledge from the codebook, derived from the labeled domain. Meanwhile, the second term ensures that the labeled domain aligns with the representations of the unlabeled domain encoded within the codebook. Additionally, α is a scalar parameter controlling the relative alignment strength. By optimizing $\mathcal{L}_1$, the codebook's latent space is structured to capture shared representations, thereby promoting effective alignment between both domains while minimizing the domain gap.

$$\mathcal{L}_1 = \alpha \cdot (\|C_i(s) - f_t\|_2) + \|C_j(t) - f_s\|_2 \tag{1}$$

3.2. Class-Level Alignment

The domain-level alignment loss ($\mathcal{L}_1$) focuses on reducing the overall domain discrepancy without explicitly considering class separation, which may lead to confusion. A bottleneck layer is introduced in the second stage to bring the same class from both domains together while maintaining greater inter-class separation. As illustrated in Figure 1, the selected quantized codes $C_i(s), C_j(t)$ are concatenated with their corresponding wavLM input features f_s, f_t and passed through the bottleneck layer, which consists of a linear transformation, a layer normalization component, and a non-linear activation function.

$$\mathcal{L}_2 = \beta \cdot \sum_{c=1}^{N_c} \|B_s^c - B_t^c\| \tag{2}$$

Class alignment is performed on the bottleneck output pairs of B_s^c and B_t^c, where B_s^c corresponds to the labeled domain and B_t^c represents the unlabeled domain, both belonging to the same class c. The total number of classes is denoted by N_c, while β is a scalar weighting parameter that regulates the alignment strength. Since the unlabeled domain lacks true labels, pseudo-labels are assigned to facilitate alignment. The pseudo-labels are derived from a SER model without adaptation, trained only with the labeled data (i.e., source domain). As defined in Equation 2, $\mathcal{L}_2$ minimizes the Euclidean distance between bottleneck features of the same class across both domains, reducing intra-class variability while maintaining inter-class separation. As illustrated in objective 2 of Figure 2, class alignment improves classification performance in the unlabeled domain by ensuring that class representations remain consistent across domains, thereby enhancing the effectiveness of cross-lingual adaptation.

3.3. InfoLoss

The previous stages operate within the Euclidean feature space, where the input wavLM features are quantized by mapping each frame to its nearest codebook entry. As described in objective 3 of Figure 2, this quantization process often leads to a loss of discriminative label information, thereby reducing the effectiveness of the learned representations. To mitigate quantization loss, we incorporate the *information loss minimization* (InfoLoss) [18], which operates within the posterior class label simplex rather than the Euclidean feature space.

The primary objective of InfoLoss is to optimize the codebook assignments by minimizing the *Kullback-Leibler* (KL) divergence between the posterior class distributions of feature vectors and the learned cluster class distributions. This objective ensures that the quantized representations preserve essential label information necessary for accurate classification. The InfoLoss function is formulated in Equation 3,

$$\mathcal{L}_3 = \gamma \cdot \frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{KL}}(p(y|\mathbf{f}_s) \| q(y|\mathbf{C}_i(s))) \tag{3}$$

where $p(y|\mathbf{f}_s)$ denotes the posterior probability of the true class for the labeled domain, and $q(y|\mathbf{C}_i(s))$ represents the class probability distribution associated with the selected code. The hyperparameter γ controls trade-off between label information preservation and quantization feature-space similarity.

Since InfoLoss relies on the posterior probabilities of the true class labels, it is applied exclusively to the labeled domain. Its primary objective is to minimize the distributional discrepancy between the model's predicted class distributions, derived from quantized feature representations, and the ideal class distributions based on ground-truth labels. Effectively, InfoLoss encourages the predictions obtained from the quantized feature vectors to align more closely with ground-truth labels, thereby retaining essential class-specific information to enhance the discriminative capacity of the learned embeddings.

Equation 4 describes the overall objective of the framework which includes the domain-level alignment loss ($\mathcal{L}_1$), class-level alignment loss ($\mathcal{L}_2$), InfoLoss ($\mathcal{L}_3$), and cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$) for the SER task. The cross-entropy loss is computed exclusively on the labeled domain.

$$L_{\text{total}} = L_1 + L_2 + L_3 + \delta \cdot L_{\mathrm{CE}} \tag{4}$$

4. Experimental Settings

4.1. Emotional Datasets

To rigorously evaluate the effectiveness of the proposed framework, we conduct cross-lingual experiments using datasets from English and Taiwanese-Mandarin, which have diverse phonetic and prosodic structures.

We utilize the MSP-Podcast corpus (v1.11) [11] as the labeled English dataset, and the BIIC-Podcast corpus [10] as the unlabeled Taiwanese-Mandarin dataset. Both databases comprise natural emotional speech collected from diverse audio recordings. For this study, we focus on four emotion categories: happiness, sadness, anger, and neutral state. The MSP-Podcast corpus contains 100,896 samples, partitioned into 57,230 for training, 12,521 for development, and 21,032 for testing. We utilize 37,663 samples from the BIIC-Podcast corpus as the unlabeled samples. This set is used for both training and evaluating the models. The emotional labels for these samples are never used during training.

Table 1: *F1-score comparison on the BIIC-Podcast dataset. Results marked with (*) indicate statistically significant improvements compared to settings without the symbol (two tailed t-test, p-value<0.05).*

Method	F1-score
No adaptation	0.480
Ladder network	0.486
Adversarial domain adaptation	0.503
kNN-VC	0.494
Proposed Method	0.551*

4.2. Implementation

The codebook is initialized with centroids obtained from the k-means clustering in the wavLM feature space, which is formed by merging frames from both domains. The codebook size (N) is empirically set to 1,024. The SER classifier follows a structure similar to the bottleneck layer, consisting of a single linear layer with a dropout rate of 0.5. The relative trade-off scalar constants are set as $\alpha = 10, \beta = 4, \gamma = 0.5$, and $\delta = 0.5$ to balance the loss components, ensuring equal emphasis on each objective. We evaluate performance using the F1-score, which accounts for *true positive* (TP), *false negative* (FN), and *false positive* (FP) rates to measure classification effectiveness. For performance comparison, we use three unsupervised baselines: the ladder network [20], an adversarial domain adaptation with gradient reversal [21], and kNN-VC [32].

5. Experimental Results

Table 1 presents the F1-scores of various methods evaluated on the BIIC-Podcast dataset, comparing the proposed approach with both domain adaptation baselines and the model without adaptation. The results indicate that the proposed method achieves the highest performance, with an F1-score of 0.551, representing a 14.8% improvement over the no-adaptation baseline. We conduct a two-tailed t-test to evaluate the results, asserting significance at p-value < 0.05. The statistical test shows that the proposed approach is significantly better than all other methods. Among the baselines, the adversarial domain adaptation method and kNN-VC strategy achieve an F1-score of 0.503 and 0.494, respectively, showing moderate improvement. The ladder network strategy achieves an F1-score of 0.486, failing to offer substantial performance gains in cross-lingual SER tasks. These findings highlight the effectiveness of the proposed approach in mitigating domain shifts while demonstrating the limitations of existing adaptation techniques in cross-lingual SER.

The performance gain achieved by the proposed approach can be attributed to the shared codebook quantization strategy, which discretizes continuous feature spaces into structured latent representations, ensuring better alignment across cross-lingual domains. The discretization process forces the model to focus on essential speech characteristics while ignoring domain-specific variations, leading to enhanced generalization. The quantization process significantly reduces the complexity of the model by representing both domains using only 1,024 dimensional discrete vectors. This reduction in complexity enhances computational efficiency, lowers memory requirements, and accelerates training and inference. Moreover, by constraining the feature space to a finite set of representative vectors, quantization helps mitigate overfitting, leading to improved generalization and higher accuracy, particularly in cross-domain adaptation. Additionally, a more structured representation aids in stabilizing optimization, ensuring more robust and consistent learning across diverse datasets.

Table 2: *Ablation study showing the impact of different components of the proposed approach on its performance.*

Ablation study	F1-score
No adaptation	0.480
Proposed Method	0.551
No Codebook	0.539
No Bottleneck	0.549
No Bottleneck, No InfoLoss	0.533
No Concatenation	0.551

The ablation study presented in Table 2 highlights the contributions of key components in the proposed approach. Removing the codebook (objectives 1 and 3) results in a performance drop to 0.539, indicating its crucial role in learning shared discrete representations that generalize across domains. Similarly, removing the bottleneck layer (objective 2) leads to a minor performance reduction (0.549), suggesting that while it refines feature representations, it is not the most critical component for adaptation. The combination of removing the bottleneck and InfoLoss (objectives 2 and 3) further lowers the F1-score to 0.533, demonstrating the significance of InfoLoss in preserving discriminative information lost during quantization. Interestingly, the model performs equivalently without concatenation, suggesting that the quantized code can represent domains as effectively as the continuous-space features while inherently possessing domain alignment capabilities, making it suitable for domain adaptation tasks. Moreover, the proposed approach achieves better alignment between both domains despite relying on pseudo-labels, which are derived from a model without adaptation and may contain errors. This result highlights the effectiveness of discrete representations in capturing essential characteristics. Overall, these findings validate the importance of quantization and information loss minimization in cross-domain adaptation and suggest that further improvements, such as refining pseudo-label quality, could enhance cross-lingual adaptation performance.

6. Conclusions

This paper presented an effective framework for unsupervised domain adaptation in speech emotion recognition by transforming continuous feature representations into a structured discrete space. The proposed approach leverages codebook quantization to enhance domain alignment while mitigating quantization-induced information loss through InfoLoss. The method facilitates robust knowledge transfer and improved generalization across datasets by promoting alignment between two highly linguistically dissimilar domains. Experimental results demonstrate the effectiveness of the proposed framework, achieving an F1-score of 0.551, which represents a 14.8% improvement over a model without adaptation. Additionally, the method outperforms all state-of-the-art unsupervised domain adaptation techniques, highlighting its capability to effectively bridge domain discrepancies. The ablation study further validates the importance of key components, particularly the codebook and InfoLoss, in preserving discriminative information and improving adaptation performance. As future research directions, further exploration of codebook hyperparameters can provide deeper insights into optimizing quantized representations for different linguistic domains. Additionally, improving pseudo-label accuracy can enhance class-level alignment, leading to more precise adaptation in unlabeled target domains. Extending this framework to multilingual and low-resource speech datasets could further establish its generalizability and practical applicability in real-world SER systems.

7. Acknowledgement

This work was funded by NSF under grant CNS-2016719.

8. References

[1] R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215–1227, April-June 2023.

[3] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[4] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[6] L. Goncalves, D. Robinson, E. Richerson, and C. Busso, "Bridging emotions across languages: Low rank adaptation for multilingual speech emotion recognition," in *Interspeech 2024*, Kos Island, Greece, September 2024, pp. 4688–4692.

[7] S. Upadhyay, L. Martinez-Lucas, B.-H. Su, W.-C. Lin, W.-S. Chien, Y.-T. Wu, W. Katz, C. Busso, and C.-C. Lee, "Phonetic anchor-based transfer learning to facilitate unsupervised cross-lingual speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023),*, Rhodes island, Greece, 2023.

[8] S. Upadhyay, C. Busso, and C.-C. Lee, "A layer-anchoring strategy for enhancing cross-lingual speech emotion recognition," in *Interspeech 2024*, Kos Island, Greece, September 2024, pp. 4693–4697.

[9] l. Kondratenko, N. Karpov, A. Sokolov, N. Savushkin, O. Kutuzov, and F. Minkin, "Hybrid dataset for speech emotion recognition in Russian language," in *ISCA Interspeech 2023*, Dublin, Ireland, August 2023, pp. 2958–1796.

[10] S. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A. Salman, C. Busso, and C.-C. Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023, pp. 1–8.

[11] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[12] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 1823–1827.

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

[14] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: A multi-corpus study," in *Speaker Classification II*, ser. Lecture Notes in Computer Science, C. Müller, Ed. Berlin, Germany: Springer-Verlag Berlin Heidelberg, August 2007, vol. 4441, pp. 43–56.

[15] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proc. Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*, 2011.

[16] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, July 2013.

[17] Y. Sun, W. Dong, X. Li, L. Dong, G. Shi, and X. Xie, "Transvqa: Transferable vector quantization alignment for unsupervised domain adaptation," *IEEE Transactions on Image Processing*, 2024.

[18] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 7, pp. 1294–1309, 2008.

[19] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5160–5164.

[20] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[21] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[22] M. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/502e4a16930e414107ee22b6198c578f-Paper.pdf

[23] S. Amiriparian, F. Packań, M. Gerczuk, and B. W. Schuller, "Exhubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets," *arXiv preprint arXiv:2406.10275*, 2024.

[24] W. Zehra, A. Javed, Z. Jalil, H. Khan, and T. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, August 2021.

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

[26] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer Science & Business Media, 2012, vol. 159.

[27] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[28] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[29] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 696–10 706.

[30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, October 2022.

[31] E. Fix and J. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[32] P. Mote, B. Sisman, and C. Busso, "Unsupervised domain adaptation for speech emotion recognition using K-Nearest neighbors voice conversion," in *Interspeech 2024*, Kos Island, Greece, September 2024, pp. 1045–1049.