Unsupervised Domain Adaptation for Speech Emotion Recognition using K-Nearest Neighbors Voice Conversion

Pravin Mote, Berrak Sisman, Carlos Busso

Department of Electrical and Computer Engineering, The University of Texas at Dallas

pravin.mote@utdallas.edu, berrak.sisman@utdallas.edu, busso@utdallas.edu

Abstract

Abundant speech data for *speech emotion recognition* (SER) is often unlabeled, rendering it ineffective for model training. Models trained on existing labeled datasets struggle with unlabeled data due to mismatches in data distributions. To avoid the cost of annotating speech data, it is imperative to explore unsupervised adaptation techniques to leverage the potential of unlabeled data. Motivated by this observation, we propose a novel use of *voice conversion* (VC) for SER, which effectively enhances emotion recognition performance on an unlabeled dataset. Our approach involves leveraging the simplicity and efficacy of the *k-nearest neighbor* (kNN)-based VC technique to transform speech samples from the unlabeled domain to the labeled domain. In contrast to conventional domain adaptation methods, our approach avoids re-training of a model on transformed unlabeled data. We achieve good results by testing transformed unlabeled samples on a model trained with a different labeled dataset.

Index Terms: Speech emotion recognition, domain adaptation, voice conversion, K-nearest neighbors

1. Introduction

Emotions play a key role in daily human interaction [1], regulating the conversations by changing and emphasizing the intention of the message. This important human capability is often missing in *human-computer interaction* (HCI) [2]. Therefore, an important research direction is to enable computers to automatically sense human emotions. Speech emotion recognition is an appealing solution given the ubiquitousness of speech-based devices. However, achieving high accuracy with a generalized model in real-world scenarios presents challenges due to disparities in emotional speech data across languages, accents, and cultural backgrounds. These disparities arise from emotionally biased datasets, variations in recording conditions, and contextual nuances within speech.

We note that models trained on one domain often perform poorly when tested on dissimilar domains [3]. Earlier strategies for enhancing generalization have involved various forms of feature normalization [4, 5], such as speaker normalization, and corpus normalization. Another strategy is to merge datasets [6] to diversify data distribution. Additionally, methods have been proposed for selecting crucial training samples [7] or assigning a higher weight to crucial training samples [8] that are similar to the unlabeled dataset. One potential approach is to further annotate extra data from the intended domain to address this challenge and improve cross-corpus performance by generalizing over mismatched domains. However, this method is limited by the unavailability and costliness of labeled data across domains. As a result, we focus on developing techniques to adapt models using unlabeled datasets. Recently, several studies have investigated domain adaptation strategies utilizing generative [9], adversarial [10], data divergence optimization [11–14], and reconstruction-based [15–17] approaches. While effective, these approaches require retraining or adapting the SER models. There is a need for domain adaptation strategies that can be easily implemented without the need to modify the already trained SER models.

This paper proposes a novel solution that utilizes voice conversion techniques to transform a domain with unlabeled data (referred to as *unlabeled domain*) into the domain used to train the models (referred to as *labeled domain*). The sentences in the unlabeled domain are transformed into sentences in the labeled domain, ensuring that test samples are well represented in the train set. The proposed idea mitigates domain mismatches and enhances the SER performance. This concept closely aligns with the task of voice conversion [18], where speech samples are converted from one speaker to another while retaining linguistic content integrity. We implement our idea using the K-nearest neighbors-voice conversion strategy [19], which is a recently proposed approach that achieves impressive results in VC despite its simplicity. The concept of kNN-VC involves replacing frames of the source speaker with frames having similar acoustic representations from the target speaker. Since *self-supervised learning* (SSL) representations can discriminate between phones [20, 21], similar frames are expected to match the phonetical information from the source speaker, ensuring content preservation while achieving voice conversion. Its simplicity provides us with the flexibility to accurately identify samples in the transformation with similar emotional content.

We extract feature representations from the labeled and unlabeled corpora using an SSL [21] model fine-tuned for the SER task. For each frame of the unlabeled sample, we identify the nearest frame in the matching pool of feature representation space of the labeled corpus using the kNN algorithm. Each unlabeled frame is then replaced with its closest match. We hypothesize that the identified nearest frame in the labeled domain would contain similar emotional characteristics as the samples in the unlabeled domain [22]. To further ensure emotional content preservation, the labeled corpus is divided into multiple pools, (referred to as *bins*), based on their emotional labels. The appropriate bin is selected as the matching pool for the kNN algorithm according to the pseudo-label of the unlabeled sample. Pseudo-labels are derived by testing all unlabeled samples on a model trained exclusively using the labeled corpus. When the bins are created with actual labels, we observe impressive performances across emotional attributes, which validate the proposed strategy. When the bins are created with pseudo labels, our proposed approach achieves performance gains as high as 8% over a baseline method.

The main contribution of this paper is the proposed unsupervised domain adaptation method, which improves SER performance in cross-corpus settings, eliminating the need for repetitive model training on unlabeled data. We study the utilization of the kNN-based VC in unsupervised domain adaptation for SER. Notably, our results demonstrate that dividing the entire labeled corpus into sub-pools based on emotional labels effectively enhances the ability of the kNN algorithm to select emotionally closer samples.

2. Related Work

2.1. Domain Adaptation for SER

Different approaches have been proposed to reduce the mismatch between train and test domains in SER. Motiian et al. [11] proposed a divergence-based domain adaptation method. Long et al. [12] explored statistical approaches to reduce the distance between multiple domains such as *maximum mean discrepancy* (MMD), *correlation alignment* (CORAL), and *contrastive domain discrepancy* (CCD). Rozantsev et al. [13] proposed separate architectures with a common regularizer for fine-tuning components, rather than sharing weights between labeled and unlabeled domains, to achieve domain adaptation. Gopalan et al. [14] connected samples from multiple domains using geodesic curves on the Grassmann manifold, demonstrating reduced performance gaps in cross-corpus settings. Liu et al. [9] proposed the CoGAN architecture, which employs two interconnected *generative adversarial networks* (GANs), each trained on a different domain to learn domain-invariant features. Abdelwahab and Busso [10] applied adversarial domain adaptation as an auxiliary task to learn common features across dissimilar domains by maximizing the domain classification loss. Ghifary et al. [15] and Deng et al. [16] proposed an autoencoder that minimizes a reconstruction loss across multiple domains, enabling it to learn domain-invariant features. Parthasarathy et al. [17] showcased enhanced SER performance through the utilization of ladder networks [23]. As an alternative approach, we propose a novel strategy for unsupervised domain adaptation to improve SER performance that does not require to adapt or retrain the SER model.

2.2. Leveraging speech synthesis and VC for SER

Another critical aspect of our study is the use of voice conversion in SER. Voice conversion is the process of transforming the speech of one speaker to sound like that of another speaker, while maintaining linguistic content and prosodic features. A range of strategies have been effectively utilized to achieve high-quality voice conversion, including feedforward neural networks [24], *long short-term memory* (LSTM) [25], attention-based encoder-decoder [26] and GANs [27]. Many-to-many nonparallel voice conversion techniques such as StarGAN [28], while promising, are often computationally complex. In contrast, Baas et al. [19] proposed a simpler method known as kNN-VC. This method replaces each frame from the source sample with its nearest neighbor frame from a pool of target samples. The approach achieves high-quality VC performance with nonparallel training data.

Various attempts have been made to explore the applications of speech synthesis and VC to improve SER performance. To address the data scarcity issue, Schuller et al. [29] proposed generating emotional data to enrich the training dataset for SER. Similarly, Bao et al. [30] applied cycleGAN to synthesize synthetic data similar to the target domain. Lotfian and Busso [31] introduce a framework to generate synthetic neutral utterances

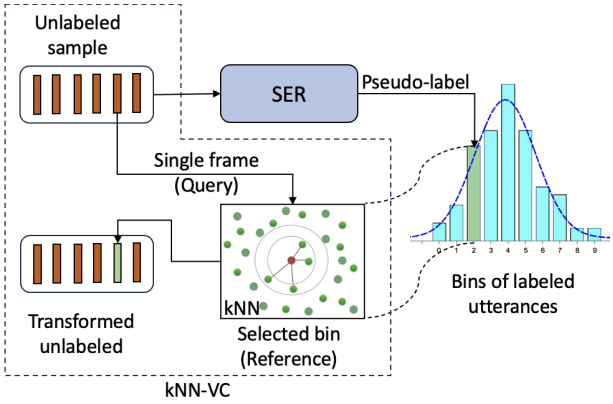

Figure 1: *Steps of the proposed domain adaptation that transforms the unlabeled domain using kNN-VC approach.*

from the target sentence without altering the lexical content. The synthetic speech was then used as a reference to assess the localized emotional modulation in the target sentence by comparing frame-by-frame their features. In this paper, we leverage voice conversion for domain adaptation to enhance SER. We specifically employ kNN-VC [19], discussed in Section 3.

3. Methodology

We envision a domain adaptation solution that does not require adapting or retraining a SER model for each new domain. We achieve this goal by transforming the unlabeled domain into sentences similar to the labeled domain, reducing the mismatches between domains. The envisioned domain transformation could be implemented with different approaches. Certain techniques such as GANs and autoencoders involve complex transformation methods that require training on both datasets, defeating the goal of preserving the original SER model. In contrast, we utilize the kNN-VC method [19], a non-parametric machine learning algorithm, to facilitate effective transformations while simplifying the overall domain adaptation process. Figure 1 presents the proposed approach. We implement our approach for the emotional attributes of arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong).

3.1. Proposed Unsupervised Domain Adaptation

The first step in our approach is to extract feature representations from the input speech (shown in red bars in Fig. 1). We use the wavLM representation [21], obtained from the HuggingFace library. We fine-tune the wavLM model using SER as the downstream task. We expect that samples with similar emotions exhibit closer proximity to each other in the feature space derived from the fine-tuned wavLM models.

We then use the kNN-VC model to transform the sentences from the unlabeled domain. Utilizing the kNN algorithm [32], unlabeled data samples are transformed into the distribution of the labeled domain projected onto the wavLM feature space. For each frame of an unlabeled sample, the k closest frames are identified from the pool of reference frames. The pool contains frames from all samples of the labeled dataset. The mean of the k closest frames replaces the corresponding frame from the unlabeled sample. We use the cosine distance to determine the closeness of frames in the wavLM feature space. Since the wavLM feature space is fine-tuned for the SER task, we expect that phonetic units with similar emotions are close to each other. After replacing all frames of the unlabeled sample with the av-

erage nearest neighbors from the labeled domain (shown with green bars in Fig. 1), the transformed sample contains data from the labeled domain with similar emotional content, thereby reducing the domain mismatch. The simplicity of the proposed approach is underscored by the fact that we do not need to retrain the model using the transformed dataset.

3.2. Use of Bins with Similar Emotional Content

The labels of the emotional dataset are annotated at the utterance level, representing the global emotion perceived across the entire sentence. However, emotions may vary at the frame level. To address frame-level fluctuations and further enhance the accuracy of selecting frames with similar emotional content, we partition the labeled dataset into multiple bins based on the scores of the attributes. This step requires to know the emotional score of the unlabeled domain. We implement two approaches. The first approach is to use the actual annotations to create the bins to study the feasibility of the proposed strategy when the bins are perfectly set. The second strategy is to use pseudo labels. Each unlabeled sample is tested on an SER model trained using the labeled dataset to obtain a pseudo-label. The bins are set using the predicted scores, partitioning the data into reference matching pools for the kNN algorithm.

4. Experimental Setting

4.1. Emotional Databases

We train and test our proposed framework using the MSP-Podcast [33] and MSP-IMPROV [34] corpora, given their mismatched data distributions. The MSP-Podcast corpus contains natural emotional speech samples from a diverse collection of podcasts, while the MSP-IMPROV corpus consists of acted dyadic interaction, making them very different from each other.

We use the MSP-Podcast (v1.11) as our labeled domain. This corpus comprises 151,654 audio samples extracted from English podcasts shared under Creative Commons licenses. The podcasts are segmented into smaller segments ranging from 2.75 seconds to 11 seconds. Speaking turns without background music, noise, or multiple speakers are considered. Given that most segments during natural conversations are emotionally neutral, the strategy automatically selected samples with machine learning algorithms that were predicted to have emotional content. All the selected segments were annotated with the emotional attributes arousal, valence, and dominance using a seven Likert-scale. Each sample was annotated by at least five raters. For a given attribute, the ground truth is the average score given by the annotators to each sentence. In release 1.11, the train set has 84,030 audio samples, the development set has 19,815 samples and the test set has 30,647 samples (we use test 1; test 2 and test 3 are not used).

We use the MSP-IMPROV corpus as our unlabeled domain. This corpus is a collection of dyadic interactions in English, recorded between 12 actors in a controlled environment. The dataset consists of 8,438 audio samples annotated for arousal, valence, and dominance by at least five annotators.

4.2. Implementation

In the initial phase, the wavLM extraction model is fine-tuned on the SER downstream task using the training set. The downstream task is an SER task. We use the same SER models for creating the pseudo labels in the unlabeled domain to determine the bins and for testing our results. The SER model is trained on the labeled dataset and consists of a feed-forward network with two fully connected hidden layers and 1 output layer. Each hidden layer applies layer normalization and dropout with a rate set to $p = 0.5$. The activation function is the *rectified linear unit* (ReLU). The *concordance correlation coefficient* (CCC) is used for the loss function and metric to evaluate the SER performance. Individual models are trained for arousal, dominance, and valence using Adam optimizer with a learning rate of 5e-5 on an NVIDIA GeForce RTX A6000 GPU. The wavLM features are extracted from both the labeled and unlabeled datasets using the same fine-tuned model.

Table 1: *Performance of the proposed approach using correct bins from the unlabeled domain. The table compares CCC values with and without domain adaptation. Results marked with (*) indicate statistically significant improvements compared to settings without the symbol (two tailed t-test, p-value<0.05).*

SER		Aro	Dom	Val
No adaptation		0.6488	0.4839	0.5251
kNN adaptation	No bins	0.6467	0.5179	0.5472
	5 bins	0.7657	0.6235	0.7411
	10 bins	0.8034*	0.6491*	0.7928*

For the kNN-VC algorithm, the extracted wavLM features in the labeled dataset are partitioned into N bins based on their corresponding labels. The scores of the emotional attributes are normalized within the range of 0 and 1, and the bins are uniformly created within this range. We conduct experiments for three cases: no bins, five bins, and ten bins. As described in the bottom part of Figure 1, the kNN algorithm is applied to each frame of an unlabeled utterance. A bin selected according to the pseudo-label serves as a matching pool for the kNN algorithm. We have chosen a value of k as 4. Once all the frames from the unlabeled utterance are replaced by their nearest neighbors, this transformed utterance is tested on the same SER model which was used to generate pseudo-labels.

5. Experimental Results

The experimental evaluation considers the performance of the proposed approach when the bins are obtained with labels from the unlabeled database (Sec. 5.1). This analysis assesses the potential of our proposed approach when the bins are perfectly obtained in the unlabeled set. Then, we present the results when the bins are estimated using pseudo labels (Sec. 5.2).

5.1. Results with Perfect Bins

We first demonstrate the adaptation results achieved by using the true labels of the unlabeled domain for bin selection. This analysis aims to showcase the potential of the proposed domain transformation using the kNN-VC algorithm. Table 1 presents the cross-corpus performance for arousal, valence, and dominance. The first row displays the CCC values for the unlabeled domain without adaptation. The subsequent three rows show the CCC values for the proposed domain transformation using different numbers of bins. The "No bins" condition indicates that the entire corpus is treated as the reference pool. The improvement in SER performance achieved using the proposed approach is truly extraordinary. Compared to the model without adaptation, there is a gain in CCC of 23.8% for arousal, 34.1% for dominance, and 51.0% for valence using 10 bins. The only case without improvement is for arousal under the "No bins" condition. There are clear improvements for all other cases.

5.2. Results with Bins formed with Pseudo Labels

Table 2 presents the results using pseudo labels to generate the bins in the unlabeled domain (i.e., unsupervised domain adaptation). Compared to the baseline without adaptation, the pro-

Table 2: *Performance of the proposed approach using bins assigned with pseudo-labels. The table compares CCC values with and without domain adaptation. Results marked with (*) indicate statistically significant improvements compared to settings without the symbol (two tailed t-test, p-value<0.05).*

SER		Aro	Dom	Val
No adaptation		0.6488	0.4839	0.5251
kNN adaptation	No bins	0.6467	0.5179	0.5472
	5 bins	0.6541	0.4907	0.5664*
	10 bins	0.6555	0.4853	0.5676*

Table 3: *Comparison of MMD values between labeled and unlabeled domains, before and after the kNN-VC transformation.*

MMD		Aro	Dom	Val
No adaptation		1.993	1.993	1.993
kNN adaptation	No bins	1.652	1.652	1.652
	5 bins	1.725	1.817	1.657
	10 bins	1.784	1.888	1.660

posed strategy implemented with 10 bins achieves a relative improvement of 1.0% for arousal, and 8.1% for valence. For dominance, the relative gain in CCC is 7.0% for the condition "No bins." These results indicate the potential of the proposed strategy to enhance cross-corpus SER performance without labeled data for the new domain.

Table 3 presents the *maximum mean discrepancy* (MMD) between the labeled and unlabeled domain. MMD quantifies the difference between two probability distributions, with smaller MMD values indicating greater similarity between the distributions. For each emotional attribute, the proposed kNN-VC adaptation results in reducing the MMD compared to the no adaptation case. Higher MMD values are observed with an increased number of bins, indicating that partitioning the dataset into more bins reduces the size of the reference pool, affecting the effectiveness of the transformations. By restricting the size of the reference pool, we limit the available choices for the nearest neighbors algorithm, resulting in transformations with distributions that are further distant from the distribution of the labeled domain. However, the higher performance often observed for the "10 bins" conditions indicates that matching the emotional content in the reference pool is beneficial despite the resulting variation in the distributions.

5.3. Impact of Bin Assignment Error on CCC Performance

We rely on pseudo-labels for selecting the bins as a reference pool for the kNN-VC adaptations. However, the results in Sections 5.1 and 5.2 show that the errors in the bin assignments have impacted the performance of our proposed approach (Table 1 and Table 2). This section explores the importance of selecting the correct bin. First, we analyze the bin assignment errors in the unlabeled dataset to assess the bin assignment accuracy obtained using pseudo-labels. Figure 2 illustrates the discrepancy in bin selection arising from pseudo-labels for the experiment with 5 bins. For example, for arousal, 3,850 samples were correctly assigned and 4,048 samples were incorrectly assigned to an adjacent bin (i.e., a difference of 1 bin between predicted and correct bin assignment). The figure shows that 45.6% (arousal), 41.4% (dominance), and 41.9 % (valence) of the samples were correctly assigned.

In this experiment, we aim to demonstrate the impact of bin assignment prediction by progressively replacing pseudo-labels with true labels. We perform replacement steps ranging from 0% to 100% with a step size of 25%, employing random selection. As an example, at the 50% replacement step, we select

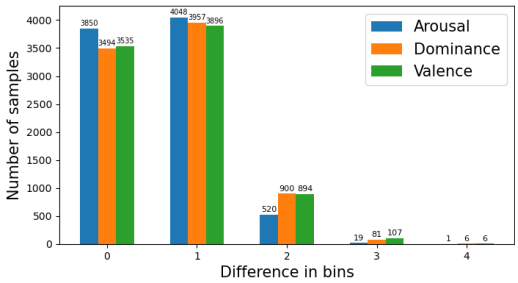

Figure 2: *Discrepancy in bin assignment for the 5-bin setting when using pseudo-labels instead of true labels (unlabeled set).*

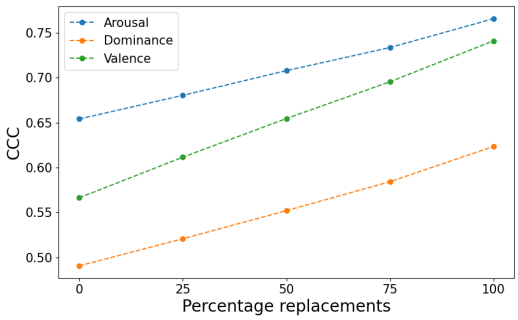

Figure 3: *Impact of bin assignment error in SER performance. The percentage replacements indicate the percentage of samples with incorrect bin assignments that are corrected. 100% represents the case when all the bin assignments are correct.*

50% of the samples from each bar of Figure 2 and substitute them with true labels. For instance, for arousal, out of the 4,048 samples with a one-bin difference, 2,024 (50%) are chosen to be corrected. Similarly, 260 out of 520 samples with a two-bin difference, and 9 out of 19 samples with a three-bin difference are selected to be corrected. 100% denotes the case where all the bin assignments are corrected. Figure 3 shows that the CCC improvement linearly increases as the bin assignment improves. This result indicates that if we can reliably assign the samples of the unlabeled domain to the correct bins, we should observe clear CCC gains.

6. Conclusions

This study proposed a novel and effective adaptation strategy using the kNN-VC algorithm. The approach transforms a new domain into the labeled domain without the need to retrain or adapt the original SER model. The approach is flexible and can be used with any SER architecture. We investigated optimizing the kNN algorithm's performance and devised a method to partition the labeled dataset into multiple bins, reducing the matching pool to sentences with similar emotions. This strategy led to improvements in SER performance for the unlabeled dataset. Notably, our algorithm enhanced CCC by 1.2% (arousal), 7.2% (dominance), and 8.2% (valence) over the SER model without the adaptation. Furthermore, we examined the impact of bin assignment errors using pseudo labels in the proposed approach, finding that improvements in the bin assignment have a direct impact on the SER performance. A future research direction involves exploring methods to enhance the pseudo-labeling accuracy. We will also evaluate this approach in other databases, especially in cross-lingual SER evaluations.

7. Acknowledgements

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

8. References

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

[2] 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.

[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, Nov. 2013, pp. 110–127.

[4] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, July-Dec 2010.

[5] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, October-December 2013.

[6] 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.

[7] 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.

[8] 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.

[9] 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

[10] 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.

[11] S. Motiian, M. Piccirilli, D. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5715–5725.

[12] L. Mingsheng, Z. Han, W. Jianmin, and M. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2208–2217.

[13] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 801–814, 2019.

[14] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *2011 international conference on computer vision*. IEEE, 2011, pp. 999–1006.

[15] M. Ghifary, B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 597–613.

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

[17] 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.

[18] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 132–157, Nov. 2020.

[19] M. Baas, B. van Niekerk, and H. Kamper, "Voice conversion with just nearest neighbors," 2023.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, December 2020, pp. 12 449–12 460.

[21] S. Chen *et al.*, "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.

[22] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 09, pp. 10 745–10 759, sep 2023.

[23] A. Rasmusi, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[24] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 954 – 964, 08 2010.

[25] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion," in *Proc. Interspeech 2016*, 2016, pp. 2453–2457.

[26] H. Kameoka, K. Tanaka, D. Kwasny, T. Kaneko, and N. Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 28, pp. 1849–1863, 2020.

[27] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[28] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[29] B. Schuller and F.Burkhardt, "Learning with synthesized speech for automatic emotion recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 5150–5153.

[30] F. Bao, M. Neumann, and N. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2828–2832.

[31] R. Lotfian and C. Busso, "Lexical dependent emotion detection using synthetic speech reference," *IEEE Access*, vol. 7, no. 1, pp. 22 071–22 085, December 2019.

[32] 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.

[33] 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.

[34] C. Busso *et al.*, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.