INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Enforcing Semantic Consistency for Cross Corpus Valence Regression from Speech using Adversarial Discrepancy Learning

Gao-Yi Chao[1,2], *Yun-Shao Lin*[1,2], *Chun-Min Chang*[1,2], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

cclee@ee.nthu.edu.tw

Abstract

Issues of mismatch between databases remain a major challenge in performing emotion recognition on target unlabeled corpus from labeled source data. While studies have shown that by means of aligning source and target data distribution to learn a common feature space can mitigate these issues partially, they neglect the effect of distortion in emotion semantics across different databases. This distortion is especially crucial when regressing higher level emotion attribute such as valence. In this work, we propose a maximum regression discrepancy (MRD) network, which enforces cross corpus semantic consistency by learning a common acoustic feature space that minimizes discrepancy on those maximally-distorted samples through adversarial training. We evaluate our framework on two large emotion corpus, the USC IEMOCAP and the MSP-IMPROV, for the task of cross corpus valence regression from speech. Our MRD demonstrates a significant 10% and 5% improvement in concordance correlation coefficients (CCC) compared to using baseline source-only methods, and we also show that it outperforms two state-of-art domain adaptation techniques. Further analysis reveals that our model is more effective in reducing semantic distortion on low valence than high valence samples.
Index Terms: valence, domain adaptation, adversarial learning, cross corpus, semantic consistency

1. Introduction

Deep learning algorithms, with its complex and non-linear learning mechanism, have brought impressive advancement to speech emotion recognition (SER) technology in recent years. While being powerful, such a data-driven learning methodology can suffer from generalizability due to the phenomenon known as dataset bias or domain shift [1]. This issue of non-robustness is especially evident when learning to perform cross corpus emotion recognition. Corpus-dependent idiosyncratic factors, such as gender distribution, languages used, recorded environments, even interaction contexts, create a situation resulting in a large mismatch between testing data (target domain) and training data (source domain). Instead of painstakingly collecting labeled data to train a predictor for each possible target scenario, domain adaption methods have been proposed to compensate for the degradation in SER performances when transferring the learned knowledge from labeled source domain to unlabeled target domain [2, 3].

Conventional SER domain adaptation approaches are based on aligning data distributions between target and source domain [4, 5]. For example, Song et al. introduced the use of maximum mean discrepancy (MMD) proposed by Borgwardt et al. in the optimization procedure of non-negative matrix factorization to address SER domain adaptation problem [6]. Other approaches have pointed to the direction that by deliberately learning an indifferentiable common feature space between source and target could mitigate domain-specific idiosyncratic factors when performing source to target emotion recognition. For example, Abdelwahab et al. used a gradient reversal layer in a three databases scenarios to predict emotion attributes of arousal, valence, and dominance [7] . Adversarial learning mechanism has also been utilized in general domain adaptation. For example, Tzeng et al. exploited GAN-based loss and untied weight sharing to reduce the difference between the source and the target [8], and Laradji et al. extended the idea by adding triplet loss and metric learning to improve the state-of-the-art unsupervised adaptation results for computer vision task [9].

The major drawback of these algorithms assumes that by aligning target and source emotion data distribution, the learned target feature representation can directly be used to transfer the source emotion label to the correct target label. However, mapping the target and source data to an indifferentiable common space does not enforce any emotion semantic consistency, i.e., source features of high valence data may be mapped to target features of low valence data. The reason for semantic distortion may be that the data distributions of the databases are completely different like the visualization presented in [7] Figure 7. In this work, our goal is to mitigate this particular issue of emotional semantic distortion in cross corpus valence regression from speech. The idea is inspired from an image classification work by Saito et al. [10]; they showed that the severity of distortion can be estimated with quantified target discrepancy, and by incorporating this discrepancy in the procedure of learning domain-indifferentiable feature space, it can improve the overall recognition performance.

Specifically, in this work, we propose a maximum regression discrepancy (MRD) network to perform cross corpus valence regression from speech. Our MRD enforces semantic consistency when learning the common acoustic feature space with adversarial discrepancy mechanism, i.e., minimizing the maximum cross corpus discrepancy. We evaluate our framework on two databases, i.e., the IEMOCAP [11] and the MSP-IMPROV [12], to perform source to target valence regression. Our methods obtains a relative gain of 10% and 5% in concordance correlation coefficients (CCC) over using source-only baseline. We compare MRD with two other domain adaptation without consistency constraint, i.e., Deep Coral (correlation alignment for deep domain adaptation) [13] and DANN (unsupervised domain adaptation by backpropogation) [14]. MRD demonstrates its superior emotion regression results over these two methods. Finally, analysis shows that MRD reduces the semantic distortion more on low valence samples than high valence samples. The rest of the paper is organized as follows: section 2 describes about database and our MRD network, section 3 presents our the experimental setup and results, and finally section 4 concludes with future work.

http://dx.doi.org/10.21437/Interspeech.2019-2037

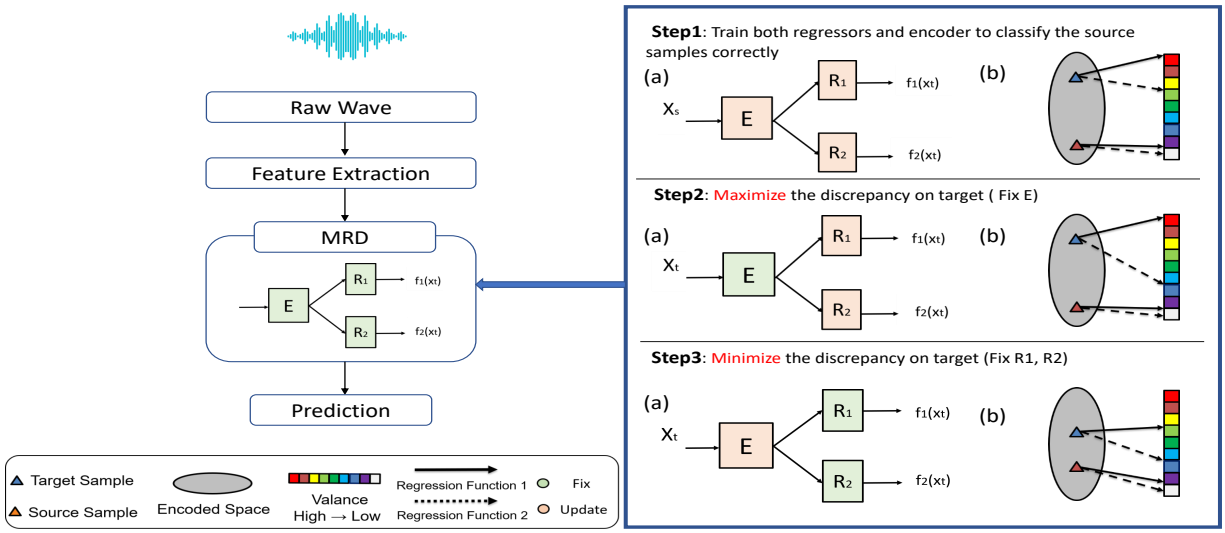

Figure 1: *Adversarial training steps of our MRD. Step1 learns two diverse valence regressors on the source data. Step2 maximizes the discrepancy by changing the regressors to detect those highly-distorted target representations. Step3 learns the encoder to minimize the discrepancy through adjusting the projected common space to reduce emotional semantic distortion. After MRD training, we finally regress the valence value of target domain sample as the average of the two regressors.*

2. Database and Method

2.1. Emotion Databases

We use two different emotion databases in our sutdy, the USC-IEMOCAP [11] and the MSP-IMPROV [12]. These two databases are one of the most commonly-used English databases in cross corpus emotion recognition research (e.g., [15, 16]). Both databases were collected in a similar setting, i.e., simulated natural dyadic interactions between actors and were labeled using similar schemes. We evaluate our valence regression experiments in a cross corpus setting, i.e., using the USC-IEMOCAP database as our source domain (training) data with the MSP-IMPROV database as our target domain (testing) data, and vice versa. Brief description of the databases is below.

2.1.1. The USC-IEMOCAP Corpus

The USC-IEMOCAP has about 12 hours of audiovisual data recorded from ten actors grouped into five sessions. The recordings were collected using scripted and improvised settings, which allowed the actors to express spontaneous emotional expressions driven by the context. The database was segmented into utterances (a total 10039 utterances). Each utterance was annotated with categorical emotion labels as well as dimensional (valence and activation) attribute with score ranges from [1,5] by at least two evaluators. The valence label used in this work is the average of the values given by the annotators.

2.1.2. The MSP-IMPROV Corpus

The MSP-IMPROV has over 9 hours of audiovisual emotion data. It consists of six dyadic sessions. In every session, two actors improvise scenarios in which one of them would utter pre-defined targeted sentences. For each of these targeted sentences, four emotional scenarios were created to contextualize the sentence to elicit happy, angry, sad and neutral. The MSP-IMPROV corpus includes the target sentences, other sentences during the improvisations, and the natural interactions between actors during the breaks. Similarly, the MSP-IMPROV was segmented into utterances (8386 utterances in total). Each utterance was annotated with categorical emotion labels as well as dimensional (valence and activation) attributes with score ranges from [1,5] using a crowdsourced evaluation scheme [17]. The valence label used in this work is the average of the values given by the annotators.

2.2. Acoustic Features

In this work, we use the IS10-paraling feature set that was used in the INTERSPEECH Paralinguistic Challenge 2010 extracted from the openSMILE toolkit [18]. This acoustic feature set consists of spectral, prosody, energy and voicing-related low level descriptors (LLDs) further processed by computing various statistical functionals (a total dimension of 1582); more detailed description can be found in the previous work [19]. We also separately z-normalize this feature set for each corpus .

2.3. Maximum Regression Discrepancy (MRD) Network

In this study, the main task is defined as a regression problem to estimate valence from speech in a cross corpus setting, i.e., source to target regression. **Figure 1** shows our entire regression framework and illustrates the adversarial discrepancy learning procedure. The MRD network is learned using labeled data, denoted as $\{X_s, Y_s\}$, from the source domain and unlabeled data, denoted as $\{X_t, Y_t\}$, from the target domain. We train a feature encoder network E that maps inputs of x_s and x_t onto a common space, and two different regressors R_1 and R_2 are trained to predict the valence labels from E-encoded feature output using labeled data. $f_1(x)$ and $f_2(x)$ is used to denote the regressed output for input x obtained by R_1 and R_2, respectively.

The key idea of our framework is based on adversarial discrepancy learning, i.e., learning an encoder E that would min-

Table 1: *Summary of our Experiment I. It lists both Pearson Correlation (PR) and Concordance Correlation (CCC) of the Source-only, Train-on-Target, DANN, Deep Coral, and our MRD model. [IEM: USC-IEMOCAP corpus, MSP: MSP-IMPROV corpus]*

	Source : IEM Target : MSP					Source : MSP Target : IEM				
	Source-Only	Train-on-Target	DANN	Deep Coral	MRD	Source-Only	Train-on-Target	DANN	Deep Coral	MRD
PR	**21.15**	42.62	22.09	25.30	**29.11**	**22.86**	43.57	22.88	24.20	**24.83**
CCC	**17.77**	38.50	21.28	23.09	**28.18**	**19.22**	39.56	21.61	14.25	**24.33**

imize the maximal semantic distortion between corpora. The two regressors, R_1 and R_2, are learned from the labeled source data and can predict source samples reliably; however the problem of domain shift will degrade the performance due to semantic distortion. The discrepancy distortion can be estimated using the inconsistency loss [20] obtained from the regressor output, $f_1(x)$ and $f_2(x)$, defined below:

$$\mathcal{L}_{dis}(X_t) = \frac{1}{K}\sum_{k=1}^{K} |f_1(x_{tk}) - f_2(x_{tk})| \tag{1}$$

K denotes the number of batches. Note that R_1 and R_2 are initialized differently, i.e., using different number of layers, to obtain diverse regressors. We then train regressors to maximize the inconsistency loss in order to identify those highly distorted samples under a fixed E. This particular learning step is important to detect those hidden distorted representations, otherwise, the two regressors tend to converge to similar outputs. Then, we finally optimize the MRD network to minimize the inconsistency loss under the same regressors, which is equivalent to narrow the distance between the similar target samples and source domain sample in the E encoded space to ensure the encoded target feature preserve the least distorted emotion information.

After the learning of MRD converges, the regressed value, r_t, for a test sample, x_t, is obtained using the following:

$$r_t = \frac{f_1(x_t) + f_2(x_t)}{2} \tag{2}$$

2.3.1. Adversarial Discrepancy Learning Procedure

The training of our cross corpus MRD network requires two regressors and one encoder to carry out the iterative adversarial discrepancy learning. We summarize the training procedure for each epoch as three major steps below (Figure 1 right):

Step 1: Both regressors and encoder are trained on the labeled source samples. The loss function used in this step is the mean square error (MSE) loss defined below:

$$\min_{E,R_1,R_2} \mathcal{L}_{mse}(X_s.Y_s) \tag{3}$$

Step 2: We update both regressors ($R1$, $R2$) and fixed the encoder E. The inverse discrepancy loss is used to increase the discrepancy to detect distorted target samples. Additionally, the source's regression loss is added to the objective function in this step. Step 2 is learned using the objective function defined as below:

$$\min_{R_1,R_2} \mathcal{L}_{mse}(X_s, Y_s) - \mathcal{L}_{dis}(X_t) \tag{4}$$

$$\mathcal{L}_{dis} = \mathbb{E}_{X_t \sim x_t}\left[|f_1(x_t) - f_2(x_t)|\right] \tag{5}$$

Step 3: We update the encoder, E, for m times to minimize the discrepancy when fixing regressors. The objective function is as follow:

$$\min_{E,R_1,R_2} \mathcal{L}_{dis}(X_t) \tag{6}$$

The hyperparameter m plays an important role in balancing the alternating min-max adversarial learning procedure between encoder and regressors. Adversarial training is often unstable, strategy based on using different learning rates [21] or applying different number of updates on generator (our encoder) and discriminator (our regressors) [22] have been used to stablize the training. In this work, we update the encoder three times ($m = 3$) in each epoch instead of just once. This parameter is determined experimentally.

3. Experimental Setup and Results

We set up two different experiments in this work. **Experiment 1** provides a comparison in the unsupervised domain adapted soeech valence regression tasks between the MSP-IMPROV and the IEMOCAP of the following models:

- **Source-only**: This is the baseline. The regressor network is trained only on the source domain and regress directly on target domain without any adaptation.
- **Train-on-Target**: This is the upper bound of the model. The regressor network is trained on the target domain and test on the target domain (no adaptation is needed) in a speaker-independent cross-validation setting.
- **DANN**: This is an unsupervised domain adaptation method through backpropagation based on method proposed by Ganin et al. [14].
- **Deep Coral**: This is a method based on correlation alignment for deep unsupervised domain adaptation proposed by Sun et al. [13].
- **MRD**: Our proposed maximum regression discrepancy network.

The parameters of the our MRD network are listed below: the number of layers for encoder and two regressors are 4, 4, and 5 respectively, and all hidden layer width is set to be 1024. We use batch normalization, dropout rates (p=0.5), activation function of SELU for all layers. The number of epochs and learning rates are determined according to different tasks. In this study, the maximum number of epochs is 100 and learning rate is chosen between the range of 5e-4 to 5e-5.

The other comparison models are implemented using similar architectures to serve as a fair comparison. Each experiment is repeated 10 times and the average of the results are reported in this work. All results are evaluated in terms of Pearsons correlation coefficient (**PR**) and concordance correlation coefficient (**CCC**) between the ground-truth labels and the regressed values. CCC is defined as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{7}$$

Experiment 2 provides a visualization on the distribution of the encoded space to analyze the generalization ability of MRD and its effectiveness as a function of the valence score.

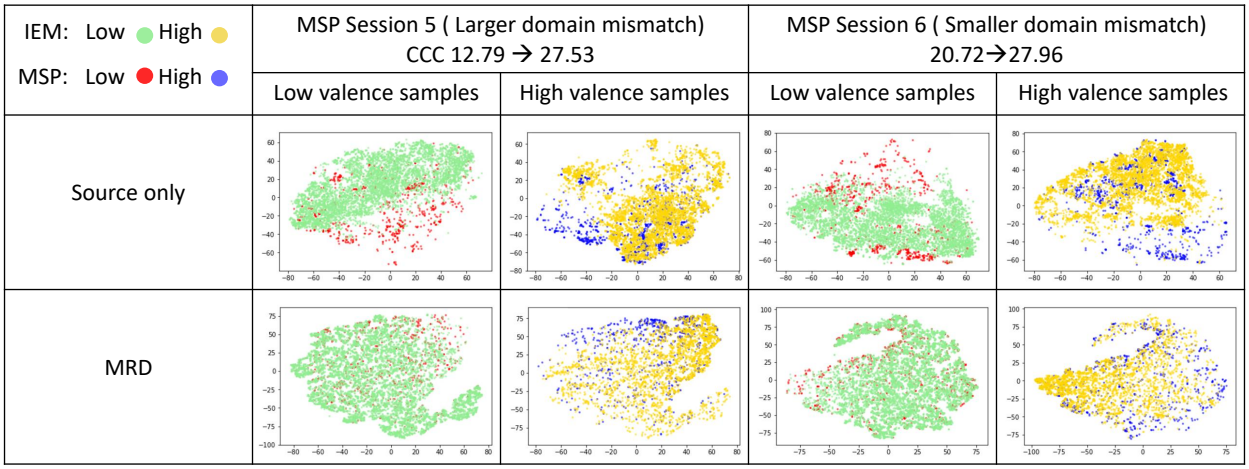

Figure 2: *Target domain: MSP-IMPROV and Source domain: IEMOCAP. We plot the 2D-projected acoustic representation in Source-only and MRD encodeed space using t-SNE. It demonstrates a reduced semantic distortion, especially more effective on the low valence samples.*

3.1. Experiment 1 Results

Table 1 summarizes our Experiment 1 results. The Train-on-Target can be seen as an upper bound of *supervised* regression performances on both the IEMOCAP and the MSP-IMPROV; it achieves 38.5% and 39.5% CCC on the two databases. Without any domain adaptation, Source-only baseline achieves 17.7% and 19.2% CCC, which indicates a severe domain mismatch that degrades the regression performances significantly. DANN improves 3.51% and 2.39% absolute over the baseline model on the MSP-IMPROV and the IEMOCAP; Deep Coral method shows an 5.32% improvement only when transferring from the IEMOCAP to the MSP-IMPROV but not vice versa. Our proposed MRD networks obtains an improvement of 10.4% and 5.01% absolute on the MSP-IMPROV and the IEMOCAP, respectively.

MRD outperforms DANN and Deep Coral due to the fact that these two other methods do not explicitly constrain the learning of the encoder when aligning different databases distributions to maintain emotion semantic consistency. Without such a consistency constraint, the encoder may generate ineffective acoustic representation because it only learns to make the two distributions ambiguous and guiding the learning to predict well only in the source domain. Valence has been stated as being a higher level affective attribute, which requires substantial contextual and cognitive appraisal [23]. It can easily lead to inconsistent semantic interpretation across domains even when features are being projected to a similar acoustic space. Furthermore, we also observe that by using domain adaptation method generally improves more when transferring from the IEMOCAP to the MSP-IMPROV, while further study is needed, we hypothesize it may due to the amount of source data available, i.e., the larger the variability exists in the database would often lead to learning a more robust representation especially when utilizing adversarial learning mechanism.

3.2. Experiment 2 Results

In Figure 2, we plot the 2D-projected acoustic representation of the Source-only model and our MRD encoder out in task of transferring from the IEMOCAP to the MSP-IMPROV using t-SNE. We plot two sessions from the MSP-IMPROV. Session 5 corresponds to the subset of MSP-IMPROV that has a lower regression performance when using Source-only model (larger domain mismatch), and Session 6 corresponds the subset that has a relatively higher regression performance prior to adaption (smaller domain mismatch).

Generally, we observe that before MRD training, the encoded feature representations from these two sessions are all very dissimilar between the two databases, and after the MRD training, the differences in the two domains representation has been decreased. By examining these t-SNE plots according to low valence and high valence samples, it is evident that our MRD help improve the low valence samples more than the high valence samples. This indicates that the semantic distortion being correctly minimized in the encoded representation from our MRD are more effective in the low valence samples. This could potentially due to simply a larger amount of low valence samples available in the source domain, or it could be related to the nature of acoustic manifestation of valence attributes, i.e., the semantic consistent space for the acoustically low valence representation may be easier to learn. However, additional investigation is needed to understand the contributing factors of emotional semantic distortion in the acoustic manifestations across different domains of emotional expressions.

4. Conclusion and Future work

In this work, we present a maximum discrepancy regression network (MRD) that learns to perform valence regression in an unsupervised cross corpus setting. Using the target sample's semantic discrepancy as feedback, our MRD adversarially learns to transform the two databases acoustic representation to a semantically-consistent and distributionally-aligned space. This particular enforcement helps improve the valence regression performances. To our knowledge, MRD exceeds the current state-of-the-art unsupervised adaptation results on regressing valence attribute. In our immediate future work, we would like to extend this framework to include lexical content, where the semantic consistency between the corpus can be further constrained by the language used. It would be interesting to further identify the underlying factors contributing to such a semantic distortion between databases especially on challenging higher level emotion attributes, such as valence, to enhance the robustness of the current emotion sensing technology.

5. References

[1] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[2] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 443–449.

[3] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

[4] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[5] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015.

[6] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[7] 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, 2018.

[8] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[9] I. Laradji and R. Babanezhad, "M-adda: Unsupervised domain adaptation with deep metric learning," *arXiv preprint arXiv:1807.02552*, 2018.

[10] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.

[11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[12] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, no. 1, pp. 67–80, 2017.

[13] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 443–450.

[14] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

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

[16] ——, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5000–5004.

[17] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, 2016.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[19] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[20] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[23] B. L. Omdahl, *Cognitive appraisal, emotion, and empathy*. Psychology Press, 2014.