Estimating Public Speaking Anxiety Using Unsupervised Transfer Learning

B Jaya Lakshmi Narayana #1, Md Mujahid #2, P Sai Lakshmi #3, M Harsha Pranathi #4, T Sai Gayathri #5

#1 Asst. Professor, Department of Computer Science and Engineering,
#2,3,4,5 Asst. Professor, Department of Computer Science and Engineering,
QIS College of Engineering & Technology

Abstract- Public speaking anxiety (PSA) ranks as a top social phobia across the world caused by various confounding factors. Motivated by the inherent data sparsity and lack of annotations in human-related applications, we propose unsupervised learning techniques to estimate PSA from speech signals. The labeled source data come from the publicly available CREMA-D dataset, while the unlabeled target data come from real-life public speaking tasks. Since fear is one of the major factors of PSA, the goal of this study is to build fear-specific representations from the labeled source data to estimate the degree of fear in the target data, and examine the extent to which the latter is associated with anxiety during the public speaking encounter. Transfer learning is performed through the domain-adversarial neural network (DANN) and Wasserstein generative adversarial network (WGAN). Results indicate that the proposed unsupervised fear specific estimates can detect public speaking anxiety with Pearson's correlation coefficient of 0.28 (p <0.01). When these fear specific estimates are combined with the degree of an individual's preparation for the public speaking task, obtained through self reports, they yield Pearson's correlation of 0.55 (p <0.01). These indicate the feasibility of leveraging labeled emotion-specific corpora for detecting human-related outcomes in real-life and provides a foundation for smart assistive technologies through the automated real-time estimation of anxiety during public speaking.

Keywords-- Public speaking anxiety, speech, unsupervised transfer learning, domain-adversarial neural network (DANN), Wasserstein generative adversarial network (WGAN)

I. INTRODUCTION

Public speaking anxiety (PSA) is a communication-based disorder that involves the experience of physiological arousal, negative cognition, and behavioral reactions in response to a real or anticipated enactment of oral presentation [1]. Previous studies in psychological and communication sciences have identified several factors contributing to PSA, including fear of not meeting the audience expectations, poor preparation and training, previous traumatic experiences, and subordinate status [2]–[4]. Such studies have mostly focused on self-reports and behavioral observations in order to estimate PSA. Physiological measures have been used for providing a complimentary view of qualitative assessments [1], [5], [6],while a limited number of studies, such as the Cicero project and Presentation Simulator, have used multimodal cues of speech intonation, facial expressions, and body gestures to model PSA [7]–[9]. Despite the encouraging results, previous methods have used in-domain data collected as part of the experimental procedures for automatically estimating PSA. Previous work has attempted transfer knowledge related to human outcomes from a source to a target domain. Recently proposed supervised approaches include progressive neural networks, which preserve the information learned from the source [10], [11]. In order to address the requirement of labeled target data, unsupervised approaches include domain adversarial neural networks (DANN) and generative adversarial neural networks (GAN) are explored and utilized to in-lab emotion datasets [12], [13]. This paper examines the feasibility of using transfer learning methods for leveraging publicly available data to estimate PSA levels using speech signals (Fig. 1). Source data come from the CREMA-D dataset [14], while target data are collected during public speaking tasks performed. Since no publicly available atasets related to speech anxiety were found and fear is a significant contributing factor to PSA [1]–[4], the proposed models will estimate the degree of fear in a given presentation using the labeled source data. We examine two unsupervised transfer learning methods, the DANN [15] and the Wasserstein generative adversarial network (WGAN) [16]. Results indicate that the proposed fear-specific estimates provided by the WGAN depict a significant Pearson's correlation of 0.28 with PSA, outperforming an in-domain model trained on the target data. We further combined the fear-specific estimates with the level of preparation and knowledge on the topic, resulting in Pearson's correlation values of 0.55 between the actual and estimated PSA values. To the best of our knowledge, no prior work has leveraged publicly available speech datasets in order to predict additional human-related outcomes in real-life. It is also the first study which attempts to perform cross-task transfer learning in affective computing applications, since it leverages emotion-specific labels to detect stress outcomes.

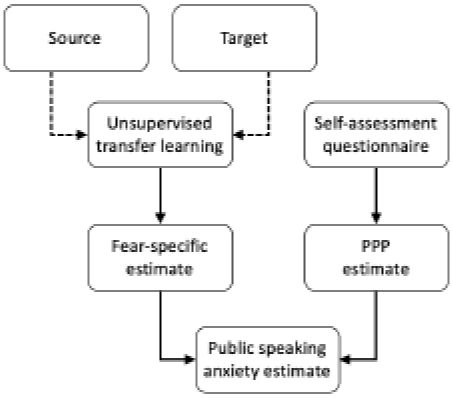

Source
Target
Unsupervised transfer learning
Self-assessment questionnaire
Fear-specific estimate
PPP estimate
Public speaking anxiety estimate

Fig. 1. Overall approach to estimating public speaking anxiety from fear specific estimates and presentation preparation performance (PPP).

II. RELATED WORKS

Transfer learning approaches, such as fine-tuning [17], and progressive neural networks [10], [11], has shown promising performance addressing the problem of sparse data or labels in the target domain. However, these methods have not fully addressed the problem of unlabeled and small-scale target data, which is a common challenge in speech-based affective computing applications. Unsupervised domain adaptation methods can leverage the above challenges by recovering similar patterns between domains and can be used for optimizing the model [18]. Ganin et al. [15] proposed the idea of adversarial learning through DANN by designing two classifiers, a domain- and a task-specific, which share a common feature extractor. Generative adversarial networks (GAN), contains a generator and a discriminator and introduced by Goodfellow et al. [19], is an appealing alternative to this task. The generator generates fake data and aims to confuse the discriminator, while the discriminator focuses on distinguishing between the real and generated data. A recently proposed modified GAN structure, the Wasserstein generative adversarial network (WGAN) [16], introduced the Wasserstein distance in the loss function, providing additional stability advantages compared to the conventional GANs. Generative and adversarial neural networks, able to fully utilize the unlabeled data, have shown promising results in the field of emotion recognition [12], [13]. In regards to public speaking, previous research has used multimodal measures of speech, body movement, and gazing, to evaluate one's overall public speaking performance and anxiety [7], [8], [20]. Previous approaches have further attempted to provide data-driven feedback for improving public speaking performance [9], [20]. Previous work has mostly focused on quantifying public speaking skills through multimodal indices (except Cicero [21]), while the estimation of public speaking anxiety has not been extensively studied using machine learning approaches. Anxiety characteristics in speech can be inherently variable for different individuals, therefore small-scale datasets might not always be able to capture the large inter-individual variability present in the entire population. Leveraging data sources from similar domains might contribute toward incorporating this variability into machine learning systems and increasing their accuracy. For this reason, this paper explores the use of out-of-domain data sources for detecting public speaking anxiety from speech signals. To the best of our knowledge, this is the first time that in-lab emotion datasets are leveraged in order to estimate various facets of affective disorder (i.e., anxiety) in spontaneous speech settings.

III. PROPOSED METHOD

The target data is collected by our group from 55 college students (32 male, 23 female) [22]. Only the data collected in a real public speaking environment with real audiences are used. The self-assessment reports filled in this study, including Trait Scale of the Communication Anxiety Inventory (CAI) [23], Personal Report of Public Speaking Anxiety (PRPSA) [24] and the Reticence Willingness to Communicate (RWTC) scale [25] are used as the evaluation standards for our systems. A custom made 6 question Presentation Preparation Performance (PPP) survey, scored in a 5-point Likert scale, is used to capture degree of preparation and knowledge on the topic. As a result, a total of 81 presentations as well as the associate self assessment reports and PPP survey is used in this study. The source data comes from the CREMA-D [14], a dataset containing 7442 clips from 91 actors and includes six emotions. Since our goal is to estimate the degree of fear in the speech signals obtained from the public speaking presentations in the target data (Section I), we included speech samples from CREMA-D belonging to neutral and fear. In order to avoid potential negative transfer that might result from the unbalanced data samples between the source and the target, we randomly selected a total of 81 clips, 41 clips belonging to fear and 40 clips belonging to neutral. These clips were converted to 16 kHz, consistently with the target data. In this section, we describe our proposed approaches, domain adversarial learning implemented with DANN (Section IV-B) and generative adversarial learning implemented with WGAN (Section IV-C), for leveraging emotion-specific knowledge in the labeled source data to provide fear-specific estimates in the unlabeled target data. We will further augment the aforementioned estimates with one's degree of preparation and knowledge on the presentation topic to provide better estimations of PSA. As baseline methods, we use an in-domain learning (IDL) paradigm, in which only the target data were used to train a classifier, and out-of-domain learning (OODL), according to which a classifier was trained on the source data and was used to predict the target data (Section IV-D).

A. Audio pre-processing and feature extraction

An automated voice activity detection (VAD) is performed to detect the speech segments of the audio signals, and each audio signal corresponding to a public speaking task is separated into multiple sentences based on the VAD decision. The Interspeech 2009 Emotion Challenge feature set, commonly used in emotion recognition tasks, were extracted over each sentence, and then averaged to yield a final 192-dimensional acoustic array per public speaking session. VAD and feature extraction were performed using the OpenSmile toolbox [26].

B. Unsupervised transfer learning with DANN

The DANN model has two tasks: the primary task is to identify the two emotions (fear and neutral) based on the labeled source data, while the secondary task is to identify the domain difference by deciding whether an input sample comes from the source or the target data (Figure 2). In order to perform these two tasks, the model consists of shared layers learned common feature representations between the two tasks, and task-specific layers trained for the primary and secondary tasks. The weights of the layers corresponding to the task classifier are learned to provide discriminative representations between fear and neutral (primary task). The weights of the domain-specific layers are learned in order to provide indiscriminative representations between the source and the target data and reduce the shift between the two domains (secondary task). For this work, we included two shared layers and two layers per task with 16 nodes each, providing a good trade-off between the complexity of the model and the amount of available data, as also observed by Abdelwahab et al. [13]. The hyper-parameters used for the DANN architecture include the learning rate (0.001), dropout (0.2), and number of epochs (500). The output of the task classifier on the target data is the final decision for the model.

C. Unsupervised transfer learning with WGAN

WGAN contains a generator and a discriminator (Fig. 3). Both the generator and discriminator consist of a 4-layer ReLU multilayer perceptron (MLP) with 16 hidden nodes. The number of hidden nodes is empirically determined to provide a balance between the number of data samples and the dimensionality of our feature space. The generator takes a 16- dimensional random distribution as input, and generates 192- dimensional output data samples. The discriminator takes the real (source and target data) and generated data as the input. These are fed into a fully-connected neural network whose weights are trained to decide whether an input sample comes from the real or generated data. In this way, the weights of the fully-connected neural network are learned in order to reduce the distribution mismatch between the source and the target data. The hyper-parameters of the generator are empirically selected and include the learning rate (0.00005), the number of training epochs of the discriminator for every epoch of generator (5), the weight clamp range (0.01), and the total number of training epochs (500) [16]. As a final step, the discriminator classifier is refined in order to take into account the labels of interest from the source data. This is performed by fine-tuning the last layer of the fully-connected network using the labeled source data. More specifically, we replace the output layer with a 2-unit layer for the classification purpose between fear and neutral, and freeze the previous hidden layers. Fine-tuning has been performed using stochastic gradient decent and learning rate of 0.001.

D. Baseline

To assess whether the proposed unsupervised transfer learning methods benefit the performance of detecting public speaking anxiety, we propose two baseline methods. The first baseline is an out-of-domain (OOD) training, according to which a 3-layer feedforward neural network is trained on the source data and directly applied to the target data without any adaptation. This network has a hidden layer with 16 units and ReLU activation, and an output layer with a Sigmoid activation. Other hyper-parameters include the dropout rate (0.2), the optimizer (Adam), and the training epochs (150). The second baseline performs an in-domain training (IDT), which estimates the PSA value from the target data. Experimentation is performed using a leave-onesubject-out (LOSO) cross-validation. More specifically, samples from one participant are included in the test set, while the rest of the data samples are used for training. Repeat this process until all participants have been used in the test set. Linear regression was used as a model for this baseline method, since it outperformed a fully-connected neural network structure, potentially because of the limited amount of data and the variety of speakers.

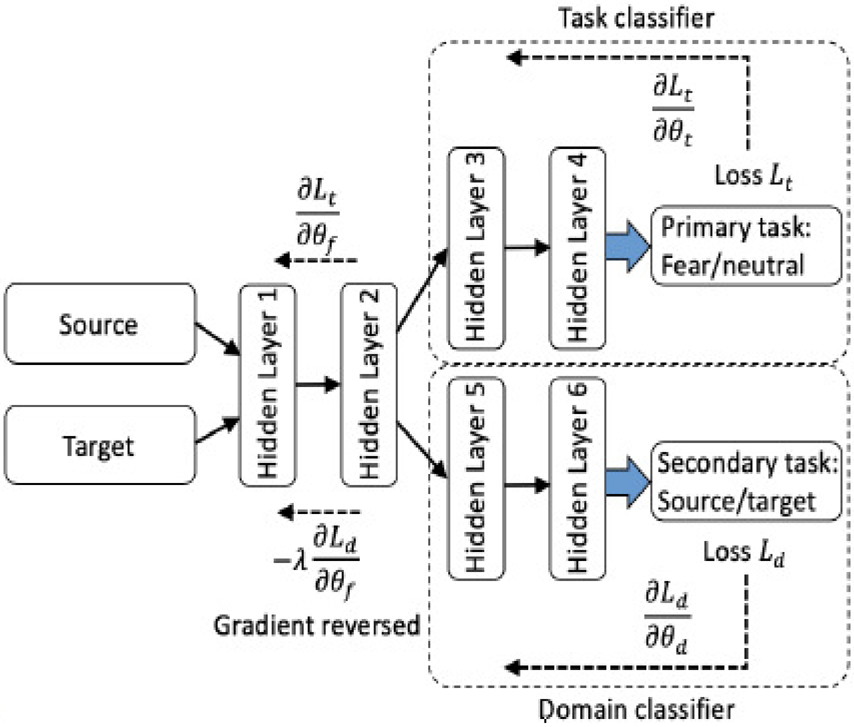

Task classifier
$\frac{\partial L_t}{\partial \theta_t}$
Loss L_t
$\frac{\partial L_t}{\partial \theta_f}$
Source
Target
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Hidden Layer 4
Primary task: Fear/neutral
Hidden Layer 5
Hidden Layer 6
Secondary task: Source/target
Loss L_d
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
Gradient reversed
$\frac{\partial L_d}{\partial \theta_d}$
Domain classifier

Fig. 2. Schematic representation of the proposed Domain Adversarial Neural Network (DANN) for unsupervised transfer learning.

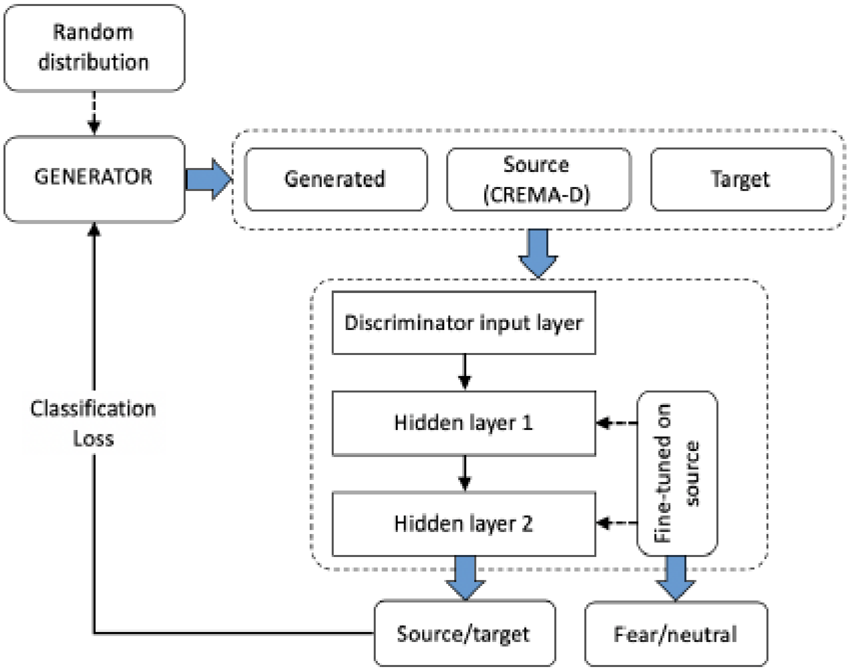


Fig. 3. Schematic representation of the proposed Wasserstein generative adversarial network (WGAN) for unsupervised transfer learning.

IV. RESULTS AND DISCUSSION

In this section, we will discuss how the output of the two proposed domain adaptation systems and the two baseline methods are used to quantify PSA (Section V-A), as well as the results obtained from the proposed systems (Section V-B).

A. Experimental settings

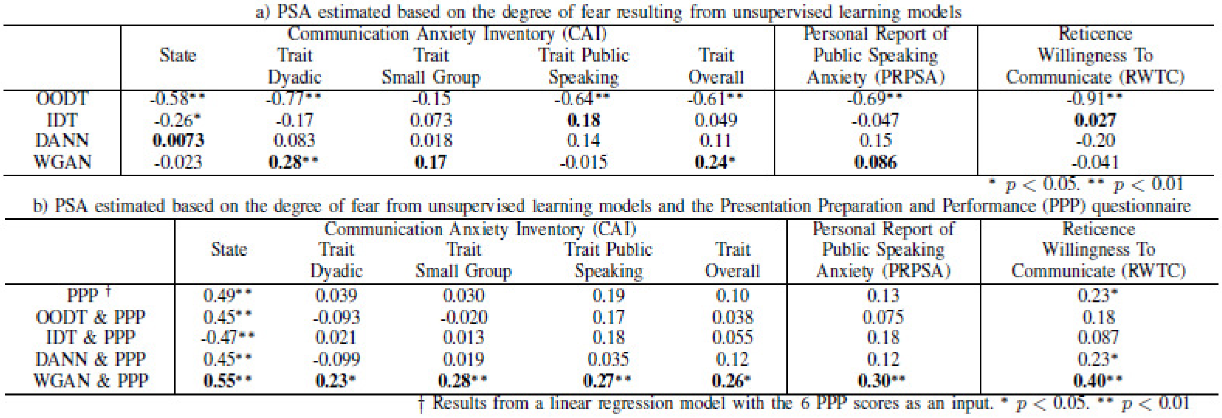
The goal of our experiments was to estimate trait- and state-based indices of PSA using the proposed unsupervised learning and baseline systems. We experimented with various PSA indices obtained from different self-reports, as discussed in Section III, including the State Scale of the CAI questionnaire, the Trait Scale of the CAI (Dyadic, Small Group, Public Speaking, and Overall constructs), as well as the PRPSA and WTC questionnaires. The DANN, WGAN, and OODT systems, which are trained based on the labeled source data of fear and neutral, provide an estimate of the amount of fear in the target audio samples. We first examine the ability of these models to learn fear-specific representations and provide reliable fear-based estimates, which will be used in order to examine their association with PSA, as obtained by the various self-reports. We further compare these with the IDT system, which is trained on the target data, providing a direct estimate of the degree of PSA. Taking into account that fear is not the only factor affecting PSA, we augmented the fearspecific estimation resulting from the unsupervised learning system with estimates related to one's level of preparation and knowledge on the presented topic, as obtained from the 6-item PPP scale. This 7-dimensional feature vector is the input of the XGBoost regression algorithm, which is trained using a LOSO cross-validation in order to yield a final PSA estimate.

B. Results

Table Ia depicts the Pearson's correlation coefficients between the fear-based estimation provided by the OODL, DANN, and WGAN systems and the PSA scores, as well as the same correlation between the actual and predicted PSA resulting from the IDL. The supervised IDT method depicted good performance, yielding the highest Pearson's correlation values for two PSA indices (CAI Trait Public Speaking, RWTC). The OODT yielded the worst performance, indicating the high mismatch between two domains. The DANN does not seem to benefit transfer learning, potentially because the small sample size in both the target and the source datasets prevent learning transferable representations. In contrast, the proposed WGAN provides significantly higher correlations compared to the other systems for many of the trait-based anxiety indices from the CAI self-report (Dyadic, Small Group, Overall). This can be due to the WGAN's ability to generate synthetic samples that could potentially leverage the domain mismatch. Table Ib provides the Pearson's correlation results when combining the fear-based estimates from the unsupervised learning models with the PPP scores. As an additional baseline, the same Table further shows the results of predicting the PSA indices solely based on the PPP questionnaire. The fearbased estimation of the WGAN combined with the PPP yields the best results and indicates that the proposed unsupervised transfer learning can potentially capture the amount of fear in audio signals of the target domain, a significant factor contributing to PSA. As expected, better results are obtained when integrating the fear-based estimations with the degree of preparation and knowledge on the topic (Table Ib) as opposed to solely relying on the fear-based estimations (Table Ia), suggesting that PSA is confounded by multiple factors. Although in this case the IDT and DANN methods provided some significant results, they still lag behind the WGAN. This might be due to the fact that these two methods are limited to the small-scale target data and learn PSA-related patterns that already exist in the preparation scores. WGAN, on the other hand, considerably boosted the correlation compared to the use of the PPP score only, or the other methods combined with PPP score. A potential factor contributing to this result is the ability of WGAN to generate new data, therefore increasing the variability of the learned representations.

TABLE I

Pearson's correlation between the estimated and actual PSA values, as measured by various self-reports, using unsupervised transfer learning through the WGAN and DANN, as well as linear regression with in-domain training (IDT) and out-of-domain training (OODT).



a) PSA estimated based on the degree of fear resulting from unsupervised learning models

	Communication Anxiety Inventory (CAI)					Personal Report of Public Speaking Anxiety (PRPSA)	Reticence Willingness To Communicate (RWTC)
	State	Trait Dyadic	Trait Small Group	Trait Public Speaking	Trait Overall		
OODT	-0.58**	-0.77**	-0.15	-0.64**	-0.61**	-0.69**	-0.91**
IDT	-0.26*	-0.17	0.073	**0.18**	0.049	-0.047	**0.027**
DANN	**0.0073**	0.083	0.018	0.14	0.11	0.15	-0.20
WGAN	-0.023	**0.28****	**0.17**	-0.015	**0.24***	**0.086**	-0.041

* $p < 0.05$. ** $p < 0.01$

b) PSA estimated based on the degree of fear from unsupervised learning models and the Presentation Preparation and Performance (PPP) questionnaire

	Communication Anxiety Inventory (CAI)					Personal Report of Public Speaking Anxiety (PRPSA)	Reticence Willingness To Communicate (RWTC)
	State	Trait Dyadic	Trait Small Group	Trait Public Speaking	Trait Overall		
PPP †	0.49**	0.039	0.030	0.19	0.10	0.13	0.23*
OODT & PPP	0.45**	-0.093	-0.020	0.17	0.038	0.075	0.18
IDT & PPP	-0.47**	0.021	0.013	0.18	0.055	0.18	0.087
DANN & PPP	0.45**	-0.099	0.019	0.035	0.12	0.12	0.23*
WGAN & PPP	**0.55****	**0.23***	**0.28****	**0.27****	**0.26***	**0.30****	**0.40****

† Results from a linear regression model with the 6 PPP scores as an input. * $p < 0.05$. ** $p < 0.01$

V. FUTURE SCOPE AND CONCLUSION

We explored the feasibility of using publicly available data from speech emotional corpora for detecting anxiety during public speaking. Since fear is inherently associated with the public speaking stimuli, we built models to classify between fear and neutral using labeled source data. Unsupervised transfer learning models, DANN and WGAN, were trained to detect the degree of fear in a speech sample. Our results indicate that WGAN is more successful in this, with Pearson's correlation of 0.28. We further combined the fear-based estimation with an individual's degree of preparation and knowledge of the topic obtained via self-assessment reports. Our results indicate that merging these two factors can achieve a Pearson's correlations at 0.55. This work can lay the foundation for developing assistive AI systems that estimate public speaking anxiety in real-time and provide in-the-moment feedback and interventions. Feedback could be potentially administer relaxation (e.g., taking a deep breath) or cognitive restructuring (e.g., providing encouraging prompts) stimuli, and can immediately provide the necessary scaffolds to motivate a healthy perception of public speaking. This work comprises a major step in achieving this goal, since it leverages publicly available data in order to provide reliable PSA estimates from individuals, for which anxiety labels might not be readily available. As part of our future work, we will attempt to use other modalities, such as physiology, in order to model several additional aspects of PSA. We will also explore additional publicly available datasets from speech and other modalities, to understand how various sources of data can benefit the transfer learning performance. We will finally attempt to objectively measure the degree of preparation through physiological signals obtained during the preparation task.

REFERENCES

[1] G. D. Bodie, "A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety," Communication education, vol. 59, no. 1, pp. 70–105, 2010.

[2] S. G. Hofmann, A. Ehlers, and W. T. Roth, "Conditioning theory: a model for the etiology of public speaking anxiety?" Behaviour Research and Therapy, vol. 33, no. 5, pp. 567–571, 1995.

[3] A. M. Bippus and J. A. Daly, "What do people think causes stage fright?: Na¨ıve attributions about the reasons for public speaking anxiety," Communication Education, vol. 48, no. 1, pp. 63–72, 1999.

[4] M. J. Beatty, "Situational and predispositional correlates of public speaking anxiety," Communication Education, vol. 37, no. 1, pp. 28– 39, 1988.

[5] D. R. Bach, K. J. Friston, and R. J. Dolan, "Analytic measures for quantification of arousal from spontaneous skin conductance fluctuations," International Journal of Psychophysiology, vol. 76, no. 1, pp. 52–55, 2010.

[6] A. Schwerdtfeger, "Predicting autonomic reactivity to public speaking: don't get fixed on self-report data!" International Journal of Psychophysiology, vol. 52, no. 3, pp. 217–224, 2004.

[7] L. Chen, G. Feng, J. Joe, C. W. Leong, C. Kitchen, and C. M. Lee, "Towards automated assessment of public speaking skills using multimodal cues," in Proceedings of the 16th International Conference on Multimodal Interaction. ACM, 2014, pp. 200–203.

[8] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, "Cicero-towards a multimodal virtual audience platform for public speaking training," in International workshop on intelligent virtual agents. Springer, 2013, pp. 116–128.

[9] J. Schneider, D. B¨orner, P. Van Rosmalen, and M. Specht, "Presentation trainer, your public speaking multimodal coach," in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015, pp. 539–546.

[10] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," arXiv preprint arXiv:1606.04671, 2016.

[11] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," in Interspeech, 2017.

[12] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 2746–2750.

[13] 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.

[14] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," IEEE Transactions on Affective Computing, vol. 5, no. 4, pp. 377–390, 2014.

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

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

[17] 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.

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

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 2014, pp. 2672– 2680.

[20] M. Chollet, T. W¨ortwein, L.-P. Morency, A. Shapiro, and S. Scherer, “Exploring feedback strategies to improve public speaking: an interactive virtual audience framework,” in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2015, pp. 1143–1154.

[21] M. Chollet, T. W¨ortwein, L.-P. Morency, and S. Scherer, “A multimodal corpus for the assessment of public speaking ability and anxiety.” In LREC, 2016.

[22] M. Yadav, K. Feng, M. N. Sakib, A. Behzadan, and T. Chaspari, “Estimating public speaking anxiety from speechsignals using unsupervised transfer learning,” in 2019 Affective Computing and Intelligent Interaction Conference (ACII 2019).

[23] S. Booth-Butterfield and M. Gould, “The communication anxiety inventory: Validation of state-and context-communication apprehension,” Communication Quarterly, vol. 34, no. 2, pp. 194–205, 1986.

[24] E. M¨ortberg, M. Jansson-Fr¨ojmark, A. Pettersson, and T. Hennlid- Oredsson, “Psychometric properties of the personal report of public speaking anxiety (prpsa) in a sample of university students in sweden,” International Journal of Cognitive Therapy, vol. 11, no. 4, pp. 421–433, 2018.

[25] J. C. McCroskey, “Reliability and validity of the willingness to communicate scale,” Communication Quarterly, vol. 40, no. 1, pp. 16–25, 1992.

[26] F. Eyben, M. W¨ollmer, and B. Schuller, “OpenSMILE: the Munich versatile and fast open-source audio feature extractor,” in Proceedings of the 18th ACM International Conference on Multimedia. ACM, 2010, pp. 1459–1462.