Estimating Public Speaking Anxiety from Speech Signals Using Unsupervised Transfer Learning

Kexin Feng[†], Megha Yadav[†], Md Nazmus Sakib[§], Amir Behzadan[§], Theodora Chaspari[†]
[†]*HUman Bio-Behavioral Signals (HUBBS) Lab*
[§]*Construction Informatics and Built Environment Research (CIBER) Lab*
Texas A&M University, College Station, TX, USA
{kexin0814,me_tam_09,mnsakib,abehzadan,chaspari}@tamu.edu

***Abstract*—Public speaking anxiety (PSA) ranks as a top social phobia across the world caused by various confounding factors, such as the novelty of the public speaking stimuli, the fear towards the audience and the potential negative evaluation, and the degree of preparation or knowledge on the task. 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 labelled source data come from the publicly available Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) dataset, while the unlabelled target data come from real-world public speaking scenarios collected by our group. Since fear is one of the major factors of PSA, the domain-adversarial neural network (DANN) and Wasserstein generative adversarial network (WGAN) are proposed to learn fear-specific representations from the source data, and reduce the mismatch between the source and the target. This results in obtaining unsupervised estimates of fear in the target data. 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 labelled emotion-specific corpora for detecting human-related outcomes in real-life and highlight the efficiency of unsupervised machine learning approaches towards this task. Results from this work provide a foundation towards assistive interventions through the automated real-time estimation of anxiety during public speaking.**

***Index Terms*—Public speaking anxiety, speech, unsupervised transfer learning, domain-adversarial neural networks (DANN), Wasserstein generative adversarial network (WGAN)**

I. Introduction

Public speaking anxiety 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 [**?**]. Research suggests that anxiety during public speaking might cause individuals to earn 10% less wages compared to their counterpeers [**?**], might prevent employees for obtaining a leadership positions [**?**], and comprises a risk factor for dropping out of college [**?**]. Many efforts have been proposed–some of which have been institutionalized–to promote public speaking of college students and industry employees [**? ?**]. Yet, most of these initiatives concentrate on the public speaking performance, while institutional training mechanisms rarely take into account the various root factors of public speaking anxiety.

Previous studies in Psychological and Communication Sciences have identified several factors contributing to public speaking anxiety. These factors involve individuals' fear of negative evaluation or not meeting the audience expectations, poor preparation and training, previous traumatic experiences, and subordinate status [**? ? ?**]. Such studies have mostly focused on self-reports and behavioral observations in order to estimate one's levels of PSA. Physiological measures have been additionally used in order to provide a complimentary view of qualitative assessments [**? ? ? ?**], while a limited number of previous studies, such as the CICERO project, have used multimodal cues related to speech intonation, facial expressions, and body gesture to model PSA [**? ? ?**]. Despite their encouraging results, previous methods have used the in-domain data collected as part of the experimental procedures in order to automatically estimate PSA. Given the inherent limitations in terms of data size and availability of labels, results from the proposed automated systems tend to reach a plateau. To the best of our knowledge, there have been no studies so far, which have attempted to leverage publicly available data in order to provide more reliable estimates of PSA. In addition to that, labelled data are even harder to find when designing a real-life system that relies on the momentary estimation of PSA, highlighting the need of using existing labelled data with unsupervised learning models for bridging potential distribution mismatch and providing accurate decisions for the outcome of interest.

Previous work has attempted both supervised and unsupervised approaches in order to transfer knowledge related to human outcomes from a source to a target domain. Recently proposed supervised approaches include neural network fine-tuning, as well as more complex neural network architectures, such as progressive neural networks, which have been utilized to preserve the information learned from the source [**? ?**]. In order to address the requirement of labelled target data, unsupervised approaches include domain adversarial neural networks (DANN) and generative adversarial neural networks (GAN) [**? ?**]. Previous approaches have attempted to transfer knowledge between emotion recognition datasets, the majority of which has been collected in in-lab settings. 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.

This paper examines the feasibility of using transfer learning methods for leveraging publicly available data to estimate PSA levels. Our work will focus on speech signals, due to the high availability of public datasets including emotional speech. Source data come from the publicly available CREMA-D dataset, while target data are collected during public speaking tasks for the purpose of the current paper. Motivated by the fact that no publicly available datasets related to speech anxiety were found and since fear is a significant contributing factor of PSA [**?** **?** **?** **?**], the proposed models are designed in order to estimate the degree of fear in a given speech signal using the labelled source data. We examine two unsupervised transfer learning methods, the first based on the Wasserstein generative adversarial network (WGAN) [**?**], while the second being the DANN [**?**]. Results indicate that the proposed fear-specific estimates provided by the WGAN depict a significant Pearson's correlation of 0.28 (p <0.01) with PSA, outperforming an in-domain model trained and tested on the target data. We further examine whether the fear-specific estimates provided by the unsupervised transfer learning models can be augmented by the level of preparation and knowledge on the topic, to better estimate the degree of PSA. Our results demonstrate that such combined factors can benefit the final performance of the system, resulting in Pearson's correlation values of 0.55 (p <0.01) between the actual and estimated PSA values.

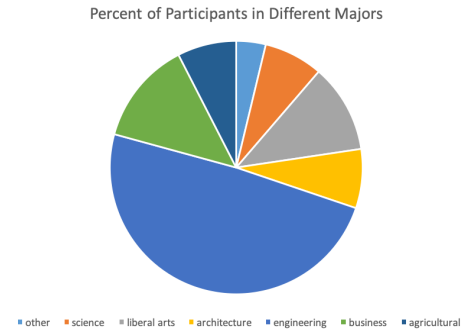
II. Previous Work

The concept of transfer learning was inspired by the ability of humans to leverage knowledge between different domains and translate this knowledge into a new domain (e.g., a child learning to recognize cars from a picture, then being able to identify cars in real life) [**?**]. In machine learning, transfer learning provides a potential solution to address the problem of sparse data and sparse (or non-existent) labels in the target domain. Previous work has used supervised and unsupervised methodologies in order to implement transfer learning models. Adaptive and incremental support vector machines (SVMs) have been explored as a solution for supervised domain adaptation in cross-corpus experiments and in order to study the potential of leveraging acted speech data in predicting emotions in spontaneous speech [**?** **?** **?**]. Sparse single-layer auto-encoders and adversarial autoencoders have been further proposed to learn class-specific low-dimensional representations from a small set of labelled data in the target domain[**?** **?**]. Neural network fine-tuning is a commonly used approach to perform transfer learning, since it relies on re-training the last layers of the source model with the target data [**?**]. Despite its intuitiveness and promising results, fine-tuning might not always be able to leverage potential mismatch between the source and the target data, and also tends to forget the weight parameters learned from the target task. a disadvantage Progressive neural networks have been also proposed as an alternative to supervised transfer learning with promising results, as they are able to leverage the representations learned from the source domain through lateral connection between a source- and a target-specific neural network [**?**].

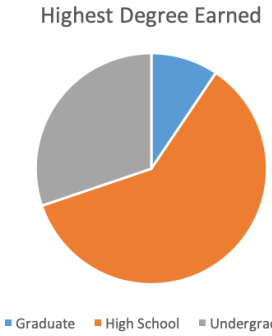
While the aforementioned work has shown promising potential of leveraging knowledge, it has not addressed the problem for the unlabeled and small-scale target data, which is a common challenge in speech-based affective computing applications. The inherent variability of speech signals resulting from the various recording settings, speaking styles, and languages can further contribute to the domain mismatch between the source and the target data. Unsupervised domain adaptation methods can leverage the above challenges by recovering similar patterns between the source and target domains and optimizing the model according to these [**?** **?**]. Ganin *et al.* [**?**] proposed the idea of adversarial learning through domain-adversarial neural networks (DANN) by designing two classifiers, a domain- and a task-specific, which share a common feature extractor. During the training process, the feature extractor aims to learn a discriminative domain-invariant feature by minimizing the error of the domain-specific classifier and maximizing the error of the task-specific classifier. This results in an adversarial behavior which learns a discriminative feature set for the task of interest which minimizes the potential domain mismatch between the source and the target data. Variants of the aforementioned adversarial learning methods have further been proposed using a modified structure of the neural network or the loss function. Generative adversarial networks (GAN) have been proposed by Goodfellow *et al.* [**?**] as an appealing alternative for this task. GAN contains two different models, a generator and a discriminator, and can be formulated as a two player game. The generator generates fake data from a random distribution and aims to confuse the discriminator, while the discriminator focuses on distinguishing between the real and generated data. In this process, both of the models can learn from each other and fully explore the patterns of data. A recently proposed modified GAN structure, the Wasserstein generative adversarial network (WGAN) [**?**], introduced the Wasserstein distance in the loss function, providing stability advantages with respect to the conventional GANs. Generative and adversarial neural networks have shown promising results in leveraging out-of-domain knowledge in the field of emotion recognition, because of their ability to fully utilize the unlabeled data. For example, Chang *et al.* [**?**] applied a modified GAN network with a convolutional structure to identify emotional valence and activation in speech. Abdelwahab & Busso [**?**] proposed a domain adversarial neural network to tackle the distribution mismatch between emotion corpora with promising results.

In regards to public speaking performance and anxiety, previous research has introduced several multimodal approaches, including information from facial expressions speech signals, and physiology. Chen *et al.*[**?**] proposed the use of audio, video, and 3D motion capturing devices to extract measures related to speech content, intonation, and body movements in order to quantify public speaking skills. In the Cicero project, Batrinca *et al.*[**?**] focused on speech, body movement, and

gaze-based measures to evaluate a speaker's overall public speaking performance. As part of the same project, Chollet *et al.* [**?** **?**] have quantified public speaking performance and anxiety through acoustic and visual modalities (e.g., speech flow, pacing, body posture) with very promising results. Aguiar *et al.* [**?**] relied on acoustic and physiological features in order to detect stress in speech during public speaking. In addition to these, providing feedback for improving public speaking performance has further been proposed by the Cicero and the Presentation Trainer systems [**?** **?**]. Feedback was provided in various visual and haptic forms with encouraging results. and was designed by Schneider *et al.* [**?**] With the exception of CICERO [**?**], previous work has mostly focused on quantifying public speaking skills through multimodal indices, while public speaking anxiety has not been extensively studied. Anxiety is an inherent psychological phenomenon affected by various confounding factors related to an individual's traits and task-specific conditions, therefore modeling anxiety is not as straightforward as modeling performance. 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 towards increasing this variability and yielding machine learning systems with increased accuracy. For this reason, this paper explores the use of out-of-domain data sources (e.g., emotional datasets) for detecting public speaking anxiety from speech signals. To the best of our knowledge, this is the first time that in-lab datasets of emotional speech are leveraged with transfer learning techniques in order to estimate various facets of human behavior (i.e., anxiety) in spontaneous speech settings. The present work will only focus on speech signals, while a multimodal transfer learning framework will be examined as part of our future work.



(a) Distribution of participants in majors



(b) Distribution of participants in degrees

Fig. 1. Demographics of study participants.

III. Data Description

Two different types of datasets are used for this work. The target data comes from a dataset of spontaneous public speaking tasks, while the source data come from a publicly available dataset of emotional speech. Emotional speech corpora provide a valuable source of data, which can be very useful for transfer learning, since they are fully annotated and contain a variety of speech data. Emotion is a fundamental component of human expression, therefore emotional speech data can be useful for modeling human-specific outcomes.

The target data was collected from 53 college students (30 male, 23 female) of an average of 22 years old. Participants pursued various educational degrees and majors (Fig. **??**). In the beginning of the user study, participants completed several self-assessment reports in order to measure their general and communication-based anxiety. These included the Trait Scale of the Communication Anxiety Inventory (CAI) [**?**], which captures trait-based indices of communication-based anxiety. Trait-CAI was scored against three separate constructs reflecting an individual's general anxiety when interacting in dyadic settings (Dyadic), when being part of a small group (Small Group), and when speaking in public (Public Speaking). A cumulative measure of the above capturing the overall trait-based communication anxiety is also being examined (Overall). In addition to these, participants were further instructed to complete the Personal Report of Public Speaking Anxiety (PRPSA) [**?**], an alternative survey which captures general public speaking anxiety, as well as the Reticence Willingness to Communicate (RWTC) scale mccroskey1992reliability, reflective a participant's reluctance to communicate. After completing these self-assessment reports, participants were given a reading material randomly selected from a pool of articles from online resources related to entertainment, well-being, health, technology, and education. Participants were asked to reading the article for 10 minutes. They were further told that they will be given 5 minutes to present the assigned article in front of an audience. Participants did not have to prepare any slides for the presentation, they were rather asked to remember the main points of the article and convey them in their own words. During the presentation, a lavalier microphone was used to collect the speech data at a 16kHz sampling rate and 16-bit encoding. The audience consisted on average of 7 people, including both students and professors to simulate an interview or class environment. After the presentation, participants were asked to fill out a six-question survey which aims to evaluate the preparation level and previous knowledge of the given topic. The items of the questionnaire are shown in Table **??** and are scored in a 5-point Likert scale. The answers to these items were summed to obtain an overall score of

presentation preparation and performance (PPP). At the end of the presentation, participants further had to complete the State Scale of the Communication Anxiety Inventory (CAI) [**?**], which reflects momentary anxiety during public speaking. The data collection consists of two sessions share the same procedure, 28 participants went to both sessions, while 24 participants only went to the first one and 1 participant only went to the second one. As a result, we collected a total of 81 presentation tasks from the 53 participants with an average duration of 4 minutes.

The source data comes from CRowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [**?**], a large-scale dataset contains 7442 clips from 91 actors and includes six emotions (anger, disgust, fear, neutral, and sad). We selected speech samples belonging to neutral and fear for the source data, since these were the most relevant to the PSA task [**?**]. Since our target data is unlabeled and has only 81 speech samples, utilizing all the source data may result in negative transfer. As a result, a total number of 81 clips, consisting of 41 clips belonging to fear and 40 clips belonging to neutral, were randomly selected as the source data and were pre-processed and converted to 16 kHz sampling rate and 16 bit encoding, consistently with the target data. It is worthwhile mentioning here that although there are many publicly available datasets of emotional speech, we could not find publicly available speech data on anxiety. For this reason, we decided to model the presence or absence of fear, as an emotion significantly inherent to PSA [**?** **?** **?**].

IV. Methodology

In this section, we describe our proposed approach for leveraging emotion-specific knowledge in the labeled source data to estimate PSA in the unlabeled target data. We explore the efficiency of two methods: 1) a domain adversarial learning implemented with DANN (Section **??**) and 2) a generative adversarial learning implemented with WGAN (Section **??**). As baseline methods, we use an in-domain learning (IDL) paradigm, in which the target data were used to train and test a classifier, as well as 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 **??**). The DANN, WGAN, and OODL systems implement unsupervised learning and will be trained on the labelled source data, therefore providing an estimate of the degree fear present in a given speech signal. The IDL model is trained based on the PSA labels of the target data.

A. Audio pre-processing and feature extraction

An automated voice activity detection is performed to automatically 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 voice activity detection feedback. Feature extraction is further implemented to compute the first 12 Mel Frequency Cepstral Coefficients, root-mean-square energy, zero-crossing rate, voice probability, and fundamental frequency for each sentence. Statistical values of the aforementioned descriptors, including maximum, minimum, range, time position of the maximum and minimum, average, standard deviation, skewness, kurtosis, the 1st and 2nd order coefficient of a linear regression model fitted on the temporary measures, and the corresponding quadratic error resulting from this model, were extracted over each sentence, and then averaged to yield an overall measure per public speaking session. This results in a 192-dimensional array. Voice activity detection and feature extraction were performed using the OpenSmile toolbox [**?**].

B. Unsupervised transfer learning with the domain-adversarial neural network (DANN) based model

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 **??**). In order to perform these two tasks, the model consists of shared layers, which learn 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 in order 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 in an effort to 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 in Abdelwahab et al. [**?**] The hyper-parameters used for the DANN architecture include the learning rate (0.001), dropout

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

TABLE I
Presentation Preparation and Performance (PPP) self-assessment questionnaire.

Question	Answer items
How would you rate your level of preparation for the presentation?	Excellent, Good, Fair, Poor, Terrible
How would you rate your performance during the presentation?	Excellent, Good, Fair, Poor, Terrible
How would you rate your prior knowledge on the topic that was given to you?	Excellent, Good, Fair, Poor, Terrible
How would you rate the level of your concentration while preparing for the presentation?	Excellent, Good, Fair, Poor, Terrible
How would you rate the difficulty of the topic that was given to you to present?	Very difficult, Difficult, Normal, Easy, Very Easy
To your best estimate, for how long were you able to talk without running out of topics (or re-iterating topics)?	1min, 2min, 3min, 4 min, 5min

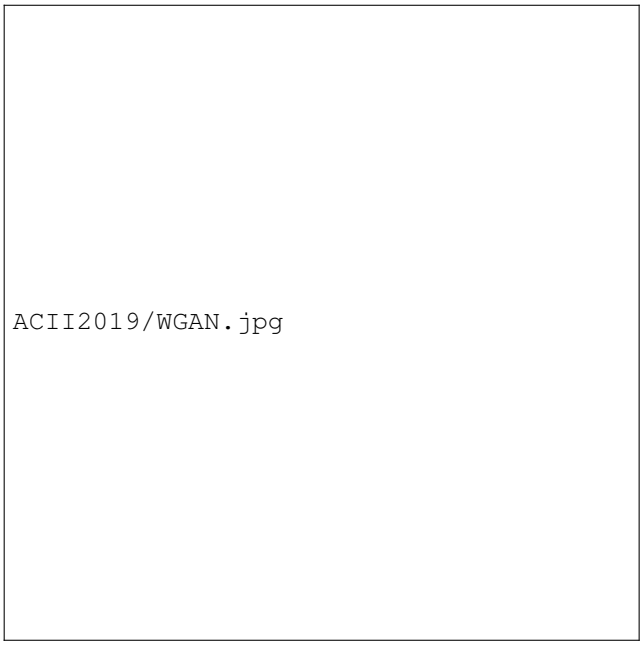


Fig. 3. Schematic representation of the proposed Wasserstein generative adversarial network (WGAN) for unsupervised transfer learning.

(0.2), and number of epochs (500). Both the primary and secondary tasks use the cross-entropy loss function. The output of the task classifier on the target data is recorded and used as the final decision for the model.

C. Unsupervised transfer learning with the Wasserstein Generative Adversarial Network (WGAN)

WGAN contains a generator and a discriminator. Both the generator and discriminator consist of a 4-layer ReLU multilayer perceptron (MLP) with 16 hidden nodes. The number of the hidden nodes is empirically determined providing 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 (Fig. **??**a). 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 (Fig. **??**b). In this way, the weights of the fulluy-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 training epochs of the discriminator for every epoch of generator (D_{iters} = 5), the weight clamp range (c = 0.01), and the total number of training epochs (500) [**?**].

As a final step, the fully-connected part of the discriminator network 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 in order to differentiate between fear and neutral using the labelled source data (Fig. **??**c). 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. In this way, the feature representations, initially learned to reduce distribution mismatch between the source and the target data, are further fine-tuned to perform a classification decision. Training of the network has been performed using stochastic gradient decent with cross entropy loss and learning rate of 0.001.

D. Baseline

In order to understand whether the proposed unsupervised transfer learning methods benefit the performance of detecting public speaking anxiety, we propose two baseline methods. An out-of-domain (OOD) training is performed during the first baseline, 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 feedforward neural network has a hidden layer with 16 units and ReLU activation, as well as an output layer with a Sigmoid activation. Other hyper-parameters include the dropout rate (0.2), the optimizer (Adam), the training epochs (150).

According to the second baseline, an in-domain training (IDT) is performed, based on which the target data are being used for both training and testing. Experimentation is performed using a leave-one-subject-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. This process is repeated for until all participants have been used in the test set. Linear regression was used as a model for both baseline methods, since it outperformed a fully-connected neural network structure, potentially because of the limited amount of data.

TABLE II

Pearson's correlation between the estimated and actual public speaking anxiety (PSA) values, as measured by various self-reports, using unsupervised transfer learning through the Wasserstein Generative Adversarial Network (WGAN) and the Domain Adversarial Neural Network (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		
Prep †	0.49**	0.039	0.030	0.19	0.10	0.13	0.23*
OODT	0.45**	-0.093	-0.020	0.17	0.038	0.075	0.18
IDT	-0.47**	0.021	0.013	0.18	0.055	0.18	0.087
DANN	0.45**	-0.099	0.019	0.035	0.12	0.12	0.23*
WGAN	**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. Experimental settings

In this section, we will discuss how we will use the output of two proposed domain adaptation systems as well as two baseline methods in order to quantify PSA (Section **??**), as well as the results obtained from the proposed systems (Section **??**).

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 coming from different self-reports, as discussed in Section **??**, 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 have been trained to perform classification between 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 are used in order to estimate the degree of PSA, as obtained by the various self-reports. Taking into account that fear is not the only factor affecting PSA, we augmented the fear-specific 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 leave-one-subject-out cross-validation in order to yield a final PSA estimate.

B. Results

Table **??**a depicts the Pearson's correlation coefficients between the fear-based estimation provided by the OODL, DANN, WGAN PSA systems and the PSA scores, as well as the same correlation between the actual and predicted PSA provided by IDL. As expected, the supervised IDT method depicted good performance, providing highest Pearson's correlation values for two PSA indices (CAI Trait Public Speaking, RWTC). The OODT yielded the worst performance, which highlights the high mismatch between the source and target domains. The DANN does not seem to benefit transfer learning, potentially because the small number of samples in both the target and the source datasets prevent this structure for 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 mismatch between the source and the target tasks.

Table **??**b provides the Pearson's correlation results when combining the fear-based estimation 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 fear-based estimation of the WGAN combined with the PPP provides the best results, outperforming the other systems across all PSA metrics. This indicates that the proposed unsupervised transfer learning can potentially capture the amount of fear in the 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 **??**b) compared to solely relying on the fear-based estimations (Table **??**a), suggesting that PSA is confounded by multiple factors. Although in this case the IDT and DANN methods provided some significant results, they still lack compared to 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, and provided the largest improvement in Pearson's correlation of about 0.26 for the CAI Trait Small Group. A potential factor attributing to this result is the ability of WGAN to generate new data, therefore increasing the variability and flexibility of the learned representations.

VI. Discussion

We explored the feasibility and efficiency of leveraging knowledge between a labeled source domain and an unlabeled target domain using several unsupervised transfer learning methods, such as the WGAN and the DANN. The WGAN architecture outperformed the DANN, as well as out-of-domain and in-domain learning across many PSA indices. Results further indicate that WGAN might be able to learn various patterns from a small amount of target data, suggesting its potential to be applied to other types of low-resource unsupervised learning problems related to human behavior.

Previous work in Psychological and Communication Sciences, which indicates that PSA is affected by a variety of factors, including one's fear of subordinate, negative evaluation or not meeting expectations, inadequate training, preparation, or knowledge on the topic, past traumatic experiences, and general trait-based anxiety. In this paper, we investigated two factors corresponding to PSA: 1) the estimated amount of fear in the speech signal, obtained through a machine learning model trained on labelled fear data and transferred to the target data through the proposed unsupervised learning approaches; and 2) the degree of preparation and knowledge on the presentation topic, obtained from a 6-point self-assessment report. Our results are in accordance to previous studies, suggesting that PSA can be better estimated when taking both these factors into account. As part of our future work, we plan to examine additional factors, such as one's trait-based anxiety and past adverse history, as well as design quantifiable indices that could potentially estimate the degree of preparation for a presentation (e.g., by studying physiological signals during preparation).

This work can provide the foundation towards developing assistive artificial intelligence systems that can estimate or predict public speaking anxiety in real-time and provide in-the-moment feedback and interventions. Feedback could come in the form of visual or haptic support and could potentially administer relaxation (e.g., taking a deep breath) or cognitive restructuring (e.g., providing encouraging prompts) stimuli. Leveraging real-time information from wearable devices, in-the-moment feedback can immediately provide the necessary scaffolds and prompts to encourage and motivate the adoption of a healthy perception of public speaking. Towards this goal, this work provides a fundamental contribution, since it aims to leverage publicly available data in order to provide reliable PSA estimates from individuals, for which anxiety labels might not be readily available.

VII. Conclusions

We explored the feasibility of leveraging 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 that can classify between fear and neutral using the labelled source data. Unsupervised transfer learning models, such as the DANN and the WGAN, were trained in order to detect the degree of fear in a speech sample, but also to leverage potential domain mismatch between the source and the target data. Our results indicate the the WGAN is more successful for doing this, reaching Pearson's correlation of 0.28 (p <0.01). We further combined the fear-based estimation yielding from the unlabelled target speech data, which were fed into the unsupervised transfer learning models, with a 6-dimensional feature vector reflecting and individual's degree of preparation and knowledge on the topic, obtained via self-assessment reports. Our results indicate that merging these two factors related to one's fear of the public speaking stimuli and degree of preparation, we are able to better estimate PSA, with Pearson's correlations reaching 0.55 (p <0.01).

As part of our future work, we will attempt to use additional modalities, such as physiology, in order to model several additional aspects of PSA. We will also explore additional publicly available datasets, in order to understand how various sources of data can benefit the transfer learning performance. We will finally attempt to objectively measure the degree of preparation, now obtained through a 6-point self-assessment scale through physiological signals obtained during the preparation task.