IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 June 2023, accepted 2 July 2023, date of publication 17 July 2023, date of current version 14 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3295832

RESEARCH ARTICLE

Multimodal Transfer Learning for Oral Presentation Assessment

SU SHWE YI TUN[1], SHOGO OKADA[1], (Member, IEEE), HUNG-HSUAN HUANG[2], AND CHEE WEE LEONG[3]

[1]Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1211, Japan
[2]Faculty of Informatics, the University of Fukuchiyama, Fukuchiyama, Kyoto 620-0886, Japan
[3]Educational Testing Service, Princeton, NJ 08540, USA

Corresponding author: Shogo Okada (okada-s@jaist.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 22H04860 and Grant 22H00536; in part by the Japan Science and Technology Agency (JST) Advanced Intelligence Project (AIP) Trilateral AI Research, Japan, under Grant JPMJCR20G6; and in part by the JST Moonshot Research and Development Program under Grant JPMJMS2237.

ABSTRACT Oral communication has consistently been ranked as a key skill, with 90 percent of hiring managers and 80 percent of business executives saying it is very important for college graduates to possess, according to a recent survey. Consequently, training and evaluating oral presentation skills remains a priority for educators worldwide, and there are increasing numbers of automated tools developed for providing feedback and assessment of such skills. However, modeling approaches typically require collecting large amounts of data and labels, which can be both expensive and laborious. In this paper, we explore the possibility of transfer learning between two different but related multimodal datasets to benefit the evaluation of oral presentation performance. We utilize knowledge from a job interview dataset as pretraining material and adapt the learned knowledge from the pre-trained model to a small amount of presentation data to improve the learning of the presentation assessment task. We demonstrate the efficacy of our approach, especially in improving performance for inference on small datasets (< 100 data points), and we report our findings. Moreover, we give a comparison between the proposed TL approach and a standard TL method based on a large-scale pre-trained model. Despite the simplicity of our proposed TL approach, the results show that our approach has promise in application to smaller datasets such as ours.

INDEX TERMS Presentation skills, multimodal, transfer learning.

I. INTRODUCTION

Oral presentation skills, including public speaking and business presentation skills, are required for conveying an intended message clearly from the presenter to the audience. In fact, such skills are central to many areas, such as education, business, politics, and leadership. In recent years, developments in nonverbal behavior detection, natural language, speech processing, and machine learning have all contributed to significant progress in multimodal modeling for automated feedback and modeling of presentation skills [1], [2]. According to findings in social science [3], a good presentation requires the presence and harmony of elements related to message production, linguistic articulation, and nonverbal expression. The interplay between these elements has led to a suite of presentation corpora collected from prior research, each with a different focus and emphasis on presentation constructs.

To date, many studies have focused on the automatic assessment of oral communication tasks. References [1], [2], [4] relied on using features from various modalities to develop automatic assessment models predict the scores assigned by human expert raters. Most studies to date conduct the automatic assessment of presentation skills using both verbal and nonverbal cues in the whole presentation or thin slices extracted from a video presentation using different machine learning algorithms. With a few exceptions [5], [6], most efforts so far have relied on traditional machine learning approaches, as deep learning methods often require large amounts of *labeled* data for training, which is expensive and laborious to obtain for videos.

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

Generally, this approach to data collection has many constraints. First, it is both laborious and expensive to collect a large dataset on a specific presentation setting (e.g., public speaking [1], business presentations [7]). To train such assessment and feedback models, a large amount of data is often required for the training process to achieve a more robust and accurate model. Second, the labels collected per data point are often narrowly restricted to the targeted presentation construct due to annotation costs. Evidence of this data limitation problem has been shown in other multimodal modeling settings, such as group discussion analysis [8].

To address the problem of data limitation in presentation assessment modeling, we present a transfer learning framework to improve the scoring accuracy of our target presentation corpus with small-size data. In our work, we exploit the idea of transfer learning (TL) [9] for modeling presentation skill assessment using two different but related presentation corpora. Additionally, we experiment with taking a holistic presentation as a time-series sequence by adopting sequential modeling in our presentation assessment model.

Essentially, TL is a learning strategy that focuses on knowledge storage and transference (e.g., weight parameters in a trained model) from a source data corpus (or multiple corpora) to a different target data corpus. Specifically, we are interested in conducting a variant of TL similar to [10] in which parameters are pretrained and transferred from a time-series source corpus to another time-series target corpus. In our study, we utilized two multimodal corpora, namely, a job interview dataset [11] as our source domain and an information presentation dataset [12] as our target domain. While TL for multimodal learning is not in itself a novel idea, this work makes new contributions, as highlighted below.

Neural transfer learning between two multimodal corpora with different target labels: We present a TL approach for multimodal, time-series data by utilizing two multimodal data corpora (job interview and presentation data) with completely different, disjoint sets of construct labels. Specifically, we situate our work in the context of an impressionistic scoring of a noncognitive, behavioral task (oral presentation), which is largely unexplored. The multimodal time-series TL framework we propose affords us an analysis of which modality is the most effective in the knowledge transfer process, culminating in a set of ideal conditions for maximum benefits we can recommend. We ask ourselves the first research question (RQ1), "*Does utilizing the job interview data via model pretraining and adaptation benefit oral presentation assessment?*". To answer RQ1, the effectiveness of the proposed TL method is discussed in Section VI-A.

Investigating an effective TL strategy: We investigate effective training methods for the multimodal time-series TL framework and probe the success conditions of multimodal TL for oral presentation assessment. First, we explore appropriate training methods for neural networks by utilizing two basic strategies used in network adaptation: fine-tuning all layers and fine-tuning only the last layer of the pretrained model. This comparison enables us to analyze the impact of the source domain on the transfer learning process. Second, we investigate the potential effectiveness of different source domain tasks to improve the accuracy of the target task using TL. The second research question (RQ2) is "*In neural transfer learning for oral presentation assessment, how do we fine-tune the network with pretrained model parameters, and which source domain tasks are most beneficial for TL?*". Our findings for RQ2 are provided in Section VI-B. In addition to the findings, proposed transfer learning method is applicable for few-shot learning with less than 20 samples in Section VII-B.

Comparison with a large-scale pretrained model: It is known that using pretrained models with a large-scale corpus of audio, visual and text data is a promising approach for TL. Through comparisons between the proposed TL approach and other popular TL approaches based on a large-scale pretrained model, we confirm that our simple approach is efficient and promising for improving the prediction accuracy of presentation skill scores. To address this concern, the third research question (RQ3) we ask ourselves is "*Does the proposed TL method outperform the TL method with a large-scale pretrained model in terms of the prediction accuracy of presentation skills for the target domain task?*". The comparative results are reported in Section VI-C.

The rest of this paper is organized as follows: Section II describes existing related studies in the literature. Then, in Section III, we describe the multimodal data corpus employed. Section IV describes the proposed transfer learning algorithm. Sections V and VI present the experimental settings and results, respectively, to answer our three research questions, followed by a discussion in Section VII and conclusion in Section VIII.

II. RELATED WORK

A. MULTIMODAL PRESENTATION ASSESSMENT

Communication skills are one of the deciding factors in many social situations, and related decision-making has been widely researched and studied in recent years. In the literature, many studies have focused on the training, feedback, and assessment of communication skills, including those focused on monologue scenarios, such as public speaking [1], [2], [4], business presentations [7] and social meetings [13], as well as those focused on communication skills in dyadic interaction situations, such as job interviews [11], [14], group interactions [15], [16] and human-computer interactions [17], [18], [19].

This study focuses on the modeling of a type of presentation skill. Presentation skills, such as speaking skills, are generally well studied. Many studies focus on an analysis of multimodal nonverbal behaviors in public speaking or presentation. Rosenberg and Hirschberg [20] investigated the relationship between focused words and acoustic features and how this relationship contributes to political speakers' performance. Scherer et al. [21] investigated the effectiveness of voice quality and pause timing while speaking.

Wörtwein et al. [1] presented an assessment model of public speaking skills by using multimodal ensemble tree-trained audio-visual information. Chollet and Scherer [22] annotated ratings of full videos and thin slices (short video clips) in a corpus of public speaking presentations and evaluated machine learning models for predicting thin-slice and full video ratings.

The annotations of public speaking skills modeled in [1], [23], [24], and [25] are mainly related to nonverbal skills, including eye contact, gesture usage, and voice control in a presentation. Chen et al. [4] also presented a multimodal model to predict public speaking skills using both speech content as verbal information and prosody, hand, body, and head movements as nonverbal features. Reference [2] proposed using time-series co-occurrence features of nonverbal aspects of a presentation to improve the prediction accuracy of public speaking skills. Lepp et al. [12] collected an informative oral presentation dataset and described how information from each specific modality presented to a rater affects her judgment in the assessment of presentation tasks and then investigated the automatic assessment of presentation content using modality-specific machine learning features and models.

Automatic assessment of presentation skills can be performed using both verbal and nonverbal cues extracted from a whole video presentation [26]. With a few exceptions [5], [6], most efforts so far have relied on traditional machine learning approaches, as deep learning methods often require large amounts of *labeled* data for training, which is expensive and laborious to obtain for videos. To address this problem, Yagi et al. [7] presented a domain adaptation algorithm using the instance weighting technique for verbal and nonverbal presentation skills in a business presentation setting. The instance weighting algorithm utilized in [7] is based on a linear model and is not directly applied to deep neural network architecture.

B. MULTIMODAL TRANSFER LEARNING

TL has been widely applied and used in many computer vision and natural language processing tasks due to its ability to leverage valuable knowledge and adapt it from one domain to another, which helps improve the model. The main idea of TL is to transfer knowledge learned from one domain to another and to efficiently utilize the learned knowledge to improve the performance of the target task. In other words, TL learns target tasks using the knowledge learned from the source task. Transfer learning can be classified into three main categories following the taxonomy of Pan and Yang et al. [9] and Ruder [10]: transductive transfer, in which the scenarios of the source and target domain tasks are the same and labels are only provided in the source domain; inductive transfer, in which the scenarios of the source and target domain tasks are different, with labels provided in the target domain; and unsupervised transfer, in which no labels are provided for either the source or target domain. Following Ruder [10], we can further divide inductive transfer into two learning settings: multitask learning (MTL) and sequential transfer learning (STL). The difference is that in MTL, the source and domain tasks are learned concurrently, while in STL, the two tasks are trained one after the other.

Sequential transfer learning (STL), proposed by Ruder [10], is a popular approach in recent machine learning and natural language processing tasks due to the simplified nature of the approach and the ease of distribution of a pretrained model. The most common scenario of sequential learning is the two-phase approach called "pretraining and adapting". In the pretraining phase, the model is trained with source data, and then the target data are adapted to the source model. STL is commonly used when the data for the source and target task are not available together, when the source task contains more data than the target task, or when many adaptations are needed for the target task. The use of pretrained ImageNet [27] in various tasks, such as object detection and semantic segmentation, in computer vision has become the primary approach for STL in other areas. STL has been successfully used in NLP, where pretrained large-scale models are adopted for various kinds of tasks, such as language modeling [28], multilingual cross-corpus tasks [29], and named entity recognition [30]. Howard and Ruder [31] proposed universal sequential transfer in text classification tasks, which is successful even for smaller data.

TL approaches to the emotion recognition task have been explored in [32]. Transfer learning in emotion recognition tasks is used in single-modality (text-based, speech-based, visual-based) and multimodality settings. In a modality-specific transfer, most work aims to transfer knowledge under similar conditions, such as transfer between lab-controlled or real-life data [33] or transfer between cross-cultural data. A common approach to TL in emotion recognition tasks is domain adaptation using deep learning models [34], [35], using adversarial and generative methods [36] or using a deep learning model as a feature extractor [37], [38], [39]. Recently, large pretrained deep learning models were used as feature extractors in recent emotion recognition tasks instead of using an intermediate feature extractor. A recent approach includes the extraction of facial features [39], [40] or acoustic features [37], [38] using pretrained convolutional neural network (CNN) architectures or using pretrained language models such as BERT [41] to extract linguistic features [38]. Hazarika et al. [42] proposed a sequential inductive TL approach for emotion recognition tasks in conversations, where a hierarchical dialogue model is pretrained on multiturn conversations and then the contextual parameters are transferred to a conversational emotion classifier. Siriwardhana et al. [43] explored using three self-supervised pretrained networks to jointly fine-tune an emotion recognition task. Gideon et al. [44] determined how knowledge can be transferred among three paralinguistic tasks using pretraining and fine-tuning approaches in the same domain. Although the findings of these studies for TL-based emotion recognition show that TL is an efficient approach to improving the accuracy of social signals such as emotion

TABLE 1. An SI interview question based on teamwork.

Usually, unpleasant tasks (e.g., tedious, boring, physically demanding) are shared among employees. Please tell us about a time when you thought you were being given more than your share of unpleasant tasks. What did you do? Please provide details about the background of the situation, the behaviors you carried out in response to that situation, and what the outcome was.

type, utilizing TL for communication skill assessment tasks is largely unexplored. Note that, although the data in this paper has been utilized in [45], the focus in that effort was an in-depth comparative evaluation between sequential and non-sequential models in an oral presentation assessment task. No TL-based approach was used in their experiments.

III. DATA

Two datasets, each with a different behavioral performance task, are utilized to explore the effectiveness of TL for oral presentation assessment. The first dataset comprises crowd-sourced interview videos, while the second dataset contains oral presentation videos. While distinct differences exist between the datasets, there are some commonalities between them: (1) both are monologue-based performance tasks, (2) both were collected online using web-based multimodal systems, and (3) both feature participants of diverse backgrounds in terms of race, ethnicity and L1 language, as well as variance in lighting, ambient noise and recording environments. That said, we explain the unique aspects of each dataset below. All participants provided written, informed consent to participate in the data collection process, and the study was reviewed and approved by the Institutional Review Board of Educational Testing Service.

A. JOB INTERVIEW DATA: SOURCE

As the source domain data for the presentation assessment model, we used a job interview dataset collected by Chen et al. [11]. The job interview dataset contains a total of 1891 monologue job interview videos (with a total duration of approximately 60 hours) from 256 online participants on Amazon Mechanical Turk. The task required each participant to answer 8 different behavioral structured interview (SI) questions, covering five different behavioral aspects of a typical workplace, i.e., leadership, communication, teamwork, persuasion, and negotiation. We provide an example of one such SI question in Table 1. The task workflow consisted of (i) receiving instructions on how to record responses to the questions, (ii) 1 minute of preparation time after being shown a given question, and (iii) 2 minutes of time to provide a response. Each video response was annotated on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree) by 5 experts from a major assessment institution for the "Big Five" personality traits, i.e., extroversion, agreeableness, conscientiousness, emotional stability, and openness to experiences, as well as a "hiring recommendation" score. In training the models to predict all 6 labels, and for each trait and hiring recommendation, we followed the same binary classification settings as [11], where we classified a given response's average rating for a trait/recommendation as Low or High by using the median value of all average ratings in the entire dataset as a threshold.

TABLE 2. Score distribution of the interview dataset.

Label	High	Low
Hiring Recommendation	1149	741
Agreeableness	936	954
Conscientiousness	962	928
Emotional Stability	946	944
Extraversion	965	925
Openness	936	954

B. ORAL PRESENTATION DATA: TARGET

Our target domain data come from an oral presentation dataset collected by Lepp et al. [12]. This dataset consists of 81 informative presentation videos by students from the United States discussing various aspects of applying to colleges (i.e., where, when, and how to apply). The participants were asked to give a presentation concerning information about the college preparation procedure rather than persuading the audience to apply (i.e., explaining why). The task involved (i) generating a checklist to consider when selecting and applying to colleges, (ii) 5 minutes of preparation for the presentation, and (iii) giving a 3-minute oral presentation. The oral presentation scores were annotated by using an oral communication scoring rubric. Each presentation was scored using the content dimension of the rubric with a Likert scale of 1 to 4 (1 = deficient, 2 = weak, 3 = competent, 4 = proficient). Annotation was performed by two experienced assessment developers for each of the three modalities, i.e., audio, video, and text. If there was a discrepancy in the score level of more than one point, a third rater was asked to perform the annotation. The raters provided three types of modality-based scores (audio, video and text) for each presentation. Two types of scores were defined for the presentation assessment task: (1) **Overall score**, which is the rounded median of all 3 modality scores (audio, video and text). (2) **Modality-specific score** (modality score) for each presentation, which is derived using only one modality (audio, video or text). As a walk-through example, assume that the annotations for a given presentation are 2,2 (text), 3,3 (video), and 1,3 (audio). The overall median score across all modalities (of 1,2,2,3,3,3) is 2.5, while the modality scores are 2 (text), 3 (video) and 2 (audio). Due to the small number of instances of the two lowest classes (i.e., "weak" and "deficient"), we combined them into a single class, which resulted in a three-class distribution of Low, Middle ("competent"), and High ("proficient") scores. Figure 1 shows the distributions of both the overall score and the modality score for the presentation assessment model.

C. FEATURE EXTRACTION

Multimodal feature extraction of the source and target domain datasets was performed automatically using

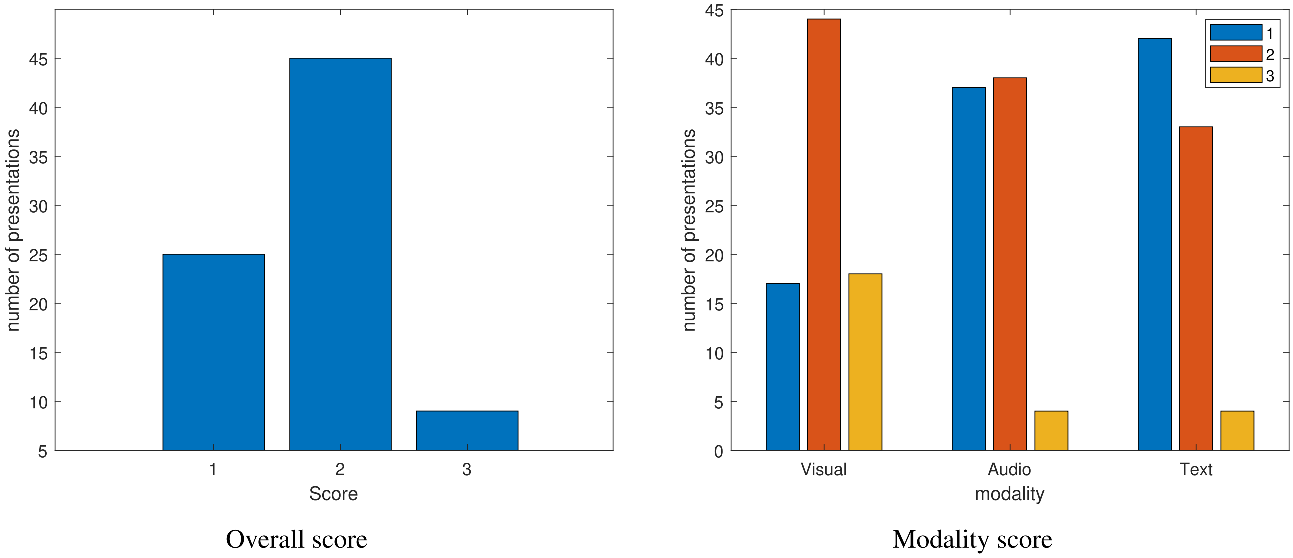


FIGURE 1. Distribution of scores by (a) overall score and (b) modality score for presentation assessment.

several feature extractors, as explained below. Specifically, we extracted acoustic information, facial expressions as a nonverbal aspect, and word-level features from the spoken utterances of the users as a verbal aspect. For the word-based features, we first performed speech-to-text conversion via a cloud-based automatic speech recognition system. Note that we validated our proposed TL model using these commonly used feature set consistent with previous research [45]. The focus of our work is on the effectiveness of our TL model by comparing it with algorithms utilizing pre-trained models trained on larger datasets than our own, rather than comparing TL models with different feature sets.

1) LINGUISTIC FEATURES

The features were extracted from the transcriptions. We extracted word embedding features for the text computed using the word2vec [46] method. First, we tokenized the words from the transcriptions and removed stop words using the Natural Language Toolkit library (NLTK) [47], and then we trained a word embedding model using the tokenized words via the Genism [48] modeling toolkit. For language model training, we used separate vocabularies for the source and target domains. The word2vec model projected our corpus with the vocabulary into the embedded vector space (embedding size of 200-D word2vec features). We converted each word in the transcription file into 200-D word2vec features and aggregated the whole word embedding from each transcription into a single embedding input.

2) ACOUSTIC FEATURES

For the audio modality, each audio file is first segmented into 5-second segments with an overlap of 1.5 seconds, and speech-based features are extracted using COVAREP [49]. The acoustic feature set contains prosodic features, voice quality information, and spectral information. Then, we compute the statistical values—mean, maximum, minimum, median, standard deviation, variance, kurtosis, skewness and percentile values—for each feature and use them as acoustic features. Last, we combine all the segments of a given audio file into one feature vector, and feature selection is performed on the targeted training dataset via a correlation matrix to select the top 100 features as the feature set for the model.

3) VISUAL FEATURES

For the video modality, each video file is first segmented into 5-second segments with an overlap of 1.5 seconds. We extract time-series features at a sampling rate of 10 FPS using the OpenFace [50] Toolkit. We then use the 2D facial landmark data of the eyes, mouth, and eyebrows to calculate the velocity and acceleration of each data point and the mean value of the 18 facial AU features. The landmarks used for each data point are described in Figure 3. Finally, we combine all the segments of a given video file into one feature vector.

Note that since the transcription is annotated with timestamps at the utterance level instead of the word level, we cannot align the audio and video frames with words in the transcriptions. Table 3 describes the details of the features used in the experiments.

IV. METHODS

In this section, we cover the notation used in our TL network architectures and the proposed network architecture for the implementation of transfer learning scenarios. Figure 2 below describes the implementation of transfer learning in our oral presentation assessment system.

A. NOTATION

We mostly follow the discussion and notation adopted in [9], [10], and [51]. We have a domain D = {X, P(X)}, where X is the feature space and P(X) is the marginal probability distribution of the feature space X. We define a task T = {Y, P(Y|X)}, where Y is the label space and P(Y|X) is the predictive objective function. Suppose that we have a source

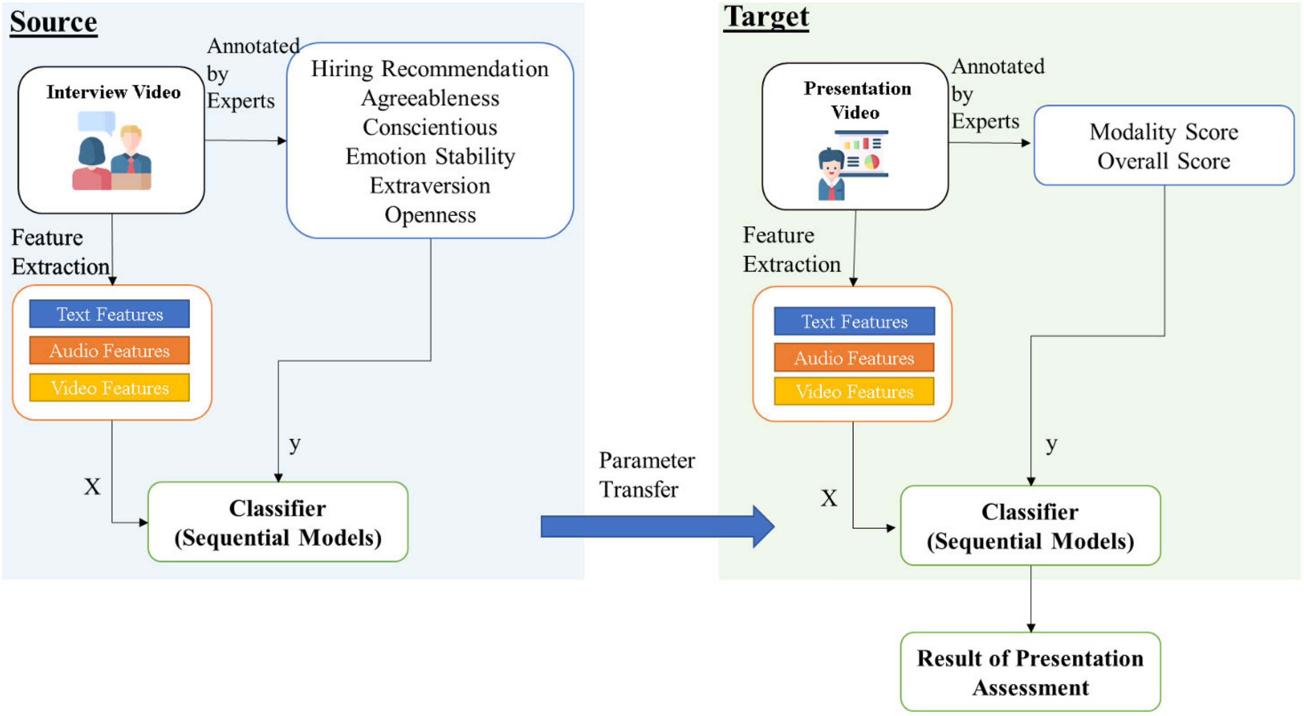


FIGURE 2. Overview of oral presentation assessment.

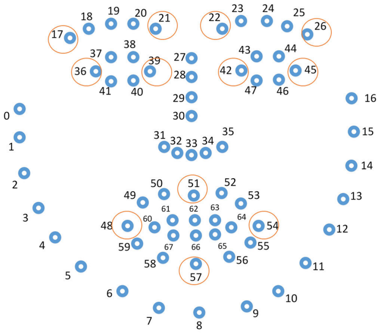


FIGURE 3. Landmarks used for extracting visual features (orange circled marks).

TABLE 3. Summary of multimodal feature sets.

Modality	Feature Names	Features
Linguistic	word2vec	200-D word2vec features
Audio	Prosodic	Fundamental Frequency (f0), voiced or unvoiced (VUV)
	Voice Quality	Normalized Amplitude Quotient (NAQ), Quasi-open Quotient (QoQ), Amplitude difference between the first two harmonics of the glottal source spectrum (H1H2), Parabolic spectral parameter (PSP), Maxima dispersion quotient (MDQ), Slope of wavelet response (peakSlope), Shape parameter of LF glottal model (Rd), Detected creaky voice (creak)
	Spectral	Mel-cepstral coefficient (MCEP 0-24), Harmonic model phase distortion mean (HMPDM 0-24), Phase distortion deviation (HMPDD 0-12)
Visual	2D facial landmarks	four points from the eyes, four points from the eyebrows, four points around the mouth
	Action Units	18 types of action units

domain D_S with a corresponding source task T_S and a target domain D_T with a corresponding task T_T. We define transfer learning as a process of improving the objective function of $P(Y|X)$ of T_T by using related information from D_S and T_S, where $D_S \neq D_T$ and $T_S \neq T_T$.

In our presentation assessment scenario, where the job interview dataset is in our source domain (D_S) and the oral presentation dataset is in our target domain (D_T), we observe the following when performing TL between the two corpora:

- $X_S \neq X_T$: the source domain is job interviews while the target domain is presentations for college preparation
- $T_S \neq T_T$: the source domain task is to predict job interview outcomes, while the target domain task is to predict oral presentation skills
- $Y_S \neq Y_T$: the source domain contains six label classes for job interviews, while the target domain contains two label classes for presentation assessment
- $P(Y_S|X_S) \neq P(Y_T|X_T)$: the outcome for the source domain is binary classification, while the target is three-class classification

Model: Suppose we have a model M that is a deep neural network model with a parameter set θ and a number of

```
Algorithm 1 TransferPretrainW
Input  : A set of parameters $\theta^{(S)} = (\theta_h^{(S)}, \theta_c^{(S)})$ of a
         pretrained source model $M^{(S)}$
Output: A set of parameters $w$
  Set $\theta_h^{(T)} \leftarrow \theta_h^{(S)}$
  Set $\theta_c^{(T)}$ to randomly initialize
  w $\leftarrow (\theta_h^{(T)}, \theta_c^{(T)})$
return w
```

layers L. The parameter set θ contains two components, θ_h and θ_c, where θ_h is the representation of each of the L hidden layers and θ_c is the representation of the output classifier.

$$M : F(x;\theta) = F_c(\ldots, F_{l_2}(F_{l_1}(x, \theta_h^{(l_1)}), \theta_h^{(l_2)}), \ldots, \theta_c) \quad (1)$$

where $F_l(.)$ is the objective output function of the l-th layer.

Parameter Sharing: Because we share the same model architecture for the source and target domains except for the output classifier, we have a target model $\mathrm{M}^{(T)}$ with parameters $\theta^{(T)} = (\theta_h^{(T)}, \theta_c^{(T)})$. Separately, a source model $\mathrm{M}^{(S)}$ is trained from scratch on a source task $\mathrm{T}^{(S)}$ with labeled data in the source domain D_S. The objective of parameter sharing is to improve $\mathrm{F}^{(T)}$ with the learned knowledge from $\mathrm{M}^{(S)}$. Our parameter sharing procedure is described in Algorithm 1.

Fine-tuning: For fine-tuning on the target model, given that we have a pretrained source model $\mathrm{M}^{(S)}$ with $\theta^{(S)}$ parameters and $L^{(S)}$ layers, the Adam [52] optimizer is updated with the learning rate in the target task so that:

$$\eta_t^{(l)} > 0, l \in [1, L^{(S)}], \eta_t^{(l)}(T) < \eta_t^{(l)}(S) \quad (2)$$

where $\eta_t^{(l)}(T)$ is the learning rate of the target model's l-th layer at iteration timestep t and the learning rate is lower than the source model learning rate.

B. PROPOSED NETWORK

Our main idea is to transfer knowledge from personality traits and hiring decision models to potentially improve the performance of the target model for predicting oral presentation performance. For this purpose, we apply TL via model pretraining and adaptation. The proposed approach is to first train a model on the job interview dataset and then fit the pretrained model to the oral presentation dataset. The notation for our proposed algorithm is described in the previous section. The main step of sequential transfer between the job interview dataset and presentation data is described below.

- **Pretraining**: Train a source model $\mathrm{M}^{(S)}$ from scratch with data in the source domain D_S
- **Parameter Transfer**: A parameter set $\theta^{(S)}$ from $\mathrm{M}^{(S)}$ is transferred to the parameters $\theta^{(T)}$ of $\mathrm{M}^{(T)}$, where $\theta_h^{(T)} = \theta_h^{(S)}$ but $\theta_c^{(T)}$ is randomly initialized.
- **Adaptation (Fine-tuning)**: Update the layers in $M^{(T)}$ with the lower learning rate from $L^{(S)}$

Next, we define the input sequence for each modality as $X_a = \{X_a^1, \ldots, X_a^n\}$, $X_v = \{X_v^1, \ldots, X_v^n\}$ and $X_t = \{X_t^1, \ldots, X_t^n\}$, which correspond to the sequences of audio, video and text data, respectively. Each input sequence data value, i.e., X_a, X_v and X_t, is mapped to the same ground-truth label Y. Each modality is trained with a separate model, such as M_a: $X_a \rightarrow$ Y, where model M is trained with audio data X_a and Y is the final output of the model. For modality fusion, late fusion is performed by concatenating all three unimodal models M_a, M_v and M_t, each of which is trained using the overall score labels.

```
Algorithm 2 Transfer Learning for Multimodal Presentation Assessment
Input  : Multimodal model for the training source
         domain: $\mathrm{M}_{mm}^{(S)}$, Multimodal training dataset
         of the source and target domain:
         $\{X^{(S),tr}, Y^{(S),tr}\}, \{X^{(T),tr}, Y^{(T),tr}\}$,
         Multimodal test dataset: $X^{te}$
Output: Model parameters: $\{\theta_{mm}\}$
Train a multimodal model as pretrained model $\mathrm{M}_{mm}^{(T)}$
  $X^{(S),tr} = \{X_a^{(S)}, X_t^{(S)}, X_v^{(S)}\}$
  $\theta_{mm}^{(S)} \leftarrow$ Pretrain($X^{(S),tr}, Y^{(S),tr}, M_{mm}^{(S)}$)
Transfer weights $\theta_{mm}^{(S)}$ of $\mathrm{M}^{(S)}$ to $\mathrm{M}^{(T)}$ (Algorithm 1)
  $w^{(T)} \leftarrow TransferPretrainW\ (\theta_{mm}^{(S)})$
Set late-fusion model with pretrained weights
  $M_{mm}^{(T)} \leftarrow F(\{x_a, x_t, x_v\}; w^{(T)})$
Fine-tune the model $M_{mm}^{(T)}$ (Equation 2)
  $X^{(T),tr} = \{X_a^{(T)}, X_t^{(T)}, X_v^{(T)}\}$
  $\theta_{mm} \leftarrow$ Fine-tune($X^{(T),tr}, Y^{(T),tr}, M_{mm}^{(T)}$)
return $\theta_{mm}$
```

Algorithm 2 shows how the transfer learning settings are applied in the oral presentation assessment task. First, we train a source model $\mathrm{M}^{(S)}$ using $X_a^{(S)}$, $X_v^{(S)}$, and $X_t^{(S)}$ from source domain D_S. Next, let $X^{(T),tr}$ denote the multimodal training dataset in our target domain D_T. The initial weights $\mathrm{w}^{(T)}$ are transferred from the pretrained model $\mathrm{M}^{(S)}$ using Algorithm 1. Fusion is performed for multimodal model $M_{mm}^{(T)}$ by concatenating the outputs of the models for each modality before final score prediction. Then, the model is fine-tuned via the backpropagation algorithm and learning rate described in Equation 2. Finally, the multimodal model is used to infer a score for each data point in the multimodal test data X^{te}.

V. EXPERIMENTS

In this section, we formulate several research questions related to the efficacy of TL by applying knowledge learned from the source domain (video interview) to the target domain (oral presentation). Specifically, as mentioned previously, we seek to answer the following questions:

- **Research Question (RQ1):** Does utilizing the job interview data via model pretraining and adaptation benefit oral presentation assessment?
- **Research Question (RQ2):** In neural transfer learning for oral presentation assessment, how do we fine-tune

the network with pretrained model parameters, and which source domain tasks are most beneficial for TL?

- **Research Question (RQ3):** Does the proposed TL method outperform the TL method with a large-scale pretrained model in terms of the prediction accuracy of presentation skills for the target domain task?

A. EXPERIMENTAL SETTINGS

1) BASELINE AND MODEL VARIANTS

Baseline: For the baseline model, we use two sequential models: LSTM and Stacked-LSTM. The LSTM model is composed of a single LSTM layer with 128 units and is used to extract the features from the input sequence data, followed by a fully connected output layer, which is used for predicting the three class labels (i.e., Low, Medium, High). The Stacked-LSTM model is composed of two LSTM hidden layers stacked together, each with 128 units, and is used to extract the features from the input sequence data, followed by a dropout layer (rate=0.5) [53]. The LSTM layers are followed by 3 time-distributed, wrapped dense layers for learning the output with 64, 32 and 16 units per layer. The time-distributed layer is flattened before the output layer, which is used for predicting the same three class labels.

Modality fusion: We opt for a late fusion method to combine all three modality outputs by concatenating all unimodal models into a fully connected layer. The concatenated layer is followed by another fully connected layer with 16 units for learning. During inferencing, this output layer generates three class labels.

Pretrained Models: For pretraining models in the source domain, the entire video interview dataset is split into training (n = 1,519) and validation (n = 371) partitions. Additionally, we ensure that no participants end up in two different partitions during the split. We use the same baseline models outlined above as our architecture for training in the source domain, except for the output layer, where a binary prediction is made for each label class. Recall that a binary classification is made for hiring recommendations and predicting the Big Five personalities. During pretraining, we first train the models using each of the six class labels. In each training loop, the model with the maximum validation accuracy is chosen as the model to be transferred.

2) MODEL HYPERPARAMETERS

We keep the features and model architectures constant except for the input layer and classifier for both the source and target domains. For supervised model pretraining, we use the Adam [52] optimizer with a learning rate of 0.001. Binary cross-entropy loss is used as the loss function, with sigmoid activation for the source domain classifier. We set the batch size to 16 and the number of epochs to 200. The model with the maximum validation accuracy is chosen as the transferred model. For source-target combination, we remove the input layer and classifier from the source domain models and add the new input layer and classifier for the target domain, i.e., oral presentation assessment. For this task, we again use the Adam [52] optimizer with a lower learning rate of 0.0001. For the target domain classifier, sparse cross-entropy loss is used as the loss function, with softmax activation. Again, we set the batch size to 16 and the number of epochs to 100. The model with the maximum validation accuracy is chosen as the best model for evaluation. For the baseline model, we use the same optimizer with a learning rate of 0.001, and the classifier is the same as that used in the target domain. The batch size is 16, and the number of epochs is 100. We use Keras with a TensorFlow backend for implementing models in both the source and target domains.

TABLE 4. Overview of large-scale data and models.

Modality	Model	Dataset
Audio	YAMNet VGGish [55]	AudioSet [54] YouTube 8-M [56]
Visual	VGG [57] MobileNet [59]	ImageNet [58] ImageNet [58]
Text	BERT [41]	Wikipedia and Books Corpus

3) EVALUATION METRIC

We use the balanced accuracy score as the main evaluation metric because the target domain dataset is highly imbalanced as a result of label bias [12]. For a more robust evaluation, we report the average balanced accuracy score for each model using 10-fold cross-validation, and data normalization is performed using Z-normalization. On the three-class classification task, the majority baseline of the balanced accuracy is 0.333.

B. COMPARISON WITH A LARGE-SCALE MODEL

We compared the proposed TL method with a standard TL method based on a large-scale pretrained model to evaluate the efficacy of the proposed method. The performance of a model trained with limited data can be improved by using a large-scale pretrained model as the feature extractor, and we used this method to validate our approach. We used pretrained CNN models to extract acoustic features and visual features and used a pretrained language model to extract linguistic features. We trained and evaluated the models on each modality. For the classification tasks, we replaced the classification (last) layer of the pretrained model with the sequential model layer mentioned in Section V-A1. Finally, we report the comparison results with the proposed TL approach. The overview of the pretrained models and the datasets used to train them are described in Table 4.

Acoustic Models We used two pretrained models for the acoustic features. First, the audio was downsampled to 16 kHz, and spectrograms were extracted as the input for a pretrained CNN to extract acoustic embedding features. Each spectrogram was computed as 64-channel, 96-time-frame log-mel-spectrogram patches (window length = 0.96 s with a hop of 0.48 s), which resulted in 2D data of size 96 × 64 for each second that were used as data points for the model. The following two CNN models were used as feature extractors:

- YAMNet: a pretrained acoustic detection model that uses the MobileNet [59] architecture and was pretrained on the AudioSet [54] dataset to predict 512 audio event classes.
- VGGish: a pretrained CNN based on the VGG [57] architecture that was pretrained on the YouTube 8-M dataset [56].

The last layer of both models was replaced with sequential models and an output layer for presentation assessment.

Visual Models We used two pretrained models for analysis of the visual features. First, each presentation video was transformed into image sequences at 10 FPS. The image sequences were then preprocessed into 224 × 224x3 patches. The pretrained CNN used the image sequences to extract visual embedding features, and then classification was performed. The following two CNN models were used as feature extractors:

- VGG: a pretrained object recognition model that contains up to 19 layers and was pretrained on the ImageNet [58] dataset.
- MobileNet: a lightweight pretrained model that uses depthwise separable convolution to reduce the model size and complexity and is mostly used in mobile and embedded vision applications.

The last layer of both models was replaced with sequential models and an output layer for presentation assessment.

Linguistic Models We used BERT [41] to extract a 768-dimensional embedding vector for each transcription. BERT is a pretrained transformer language model that has achieved state-of-the-art results in many NLP tasks. The BERT we used is uncased and was pretrained using the English Wikipedia and Books Corpus. As we describe in the feature extraction of linguistic features (Section III-C), we first tokenized each word from the transcriptions and removed stop words. Note that the maximum length of tokens that BERT can accept is 512, so we ignored subsequent words in transcripts that exceeded the maximum length. The tokens were used as input for BERT to extract 768-dimensional embedding vectors. We used the last hidden layer of BERT as the embedding for linguistic models.

VI. RESULTS

In this section, we discuss the experimental results that provide the foundation for answering the research questions posed in the previous section.

A. EFFICACY of TRANSFER LEARNING for ORAL PRESENTATION ASSESSMENT (ANSWER TO RQ1)

To answer this research question, we adopted the sequential transfer learning we proposed in Algorithm 2. First, we pretrained supervised models using the job interview dataset and then adapted (or fine-tuned) these pretrained models using the oral presentation dataset in the target domain. We evaluate the efficacy of this approach in an incremental manner, first using unimodal, then bimodal, and finally multimodal approaches.

We first perform a comparative evaluation between the LSTM and Stacked-LSTM baseline models, with the results shown in Table 5. Assuming no preference for overall or modality score labels, we evaluate both models with four criteria: (1) audio modality score, (2) video modality score, (3) text modality score, and (4) overall score prediction from the multimodal features. We also compute the average score across the four criteria for each model. On average, the LSTM model performed better than the Stacked-LSTM model, with an average balanced accuracy score of 0.462. Consequently, we selected the LSTM model as the better model for further analysis.

Using LSTM as the base model, we compare the outcomes of using a traditional supervised model and one that is enhanced with TL. Table 6 provides the results of this comparative evaluation. Table 6 compares the average balanced accuracy score between the baseline model, which is an LSTM model trained using the dataset in the target domain, and the same LSTM model enhanced with TL and fine-tuning. Except for the text modality score, all scores using TL exceed the scores of the baseline LSTM without TL.

The balanced accuracies of the unimodal approaches for modeling the overall score using LSTM are 0.350 (audio modality), 0.443 (video modality), and 0.376 (text modality), while the corresponding balanced accuracy numbers using TL are 0.481, 0.449, and 0.478, respectively. This consistency in improvement demonstrates the efficacy of TL for unimodal modeling approaches. Although bimodal approaches do not always result in an improvement for modeling the overall score (e.g., A+V (0.381) < video (0.443)), the A+V approach obtains a significant improvement when TL is performed (i.e., A+V (0.527)). This increase in balanced accuracy is again consistent across all bimodal combinations. Finally, the multimodal fusion approach (A+V+T) to modeling the overall score is again better when TL is used than when it is not, yielding an improvement of 0.066 when TL is used. For modeling the modality scores of the presentations, using TL yields an improvement of 0.025 and 0.001 for the audio and video modalities. In view of this evidence, we conclude that the utilization of knowledge from a source domain dataset can potentially lead to significant improvements in the performance of task modeling in a related but different target domain.

B. INVESTIGATING EFFECTIVE TL STRATEGIES (ANSWER TO RQ2)

To address the second research question, first, we explore appropriate fine-tuning methods in Section VI-B1. Second, we investigate the types of source domain tasks to improve the accuracy of the target task with TL in Section VI-B2.

1) COMPARISON OF FINE-TUNING METHODS

We utilized two basic strategies used in network adaptation: fine-tuning all layers and fine-tuning only the last layer of the

TABLE 5. Comparative evaluation of LSTM and Stacked-LSTM using a pretrained model in the source domain that is trained on the hiring recommendation label, applied to the target domain.

Score	LSTM	Stacked-LSTM
Audio Modality Score	**0.587**	0.535
Video Modality Score	**0.391**	0.294
Text Modality Score	0.436	**0.538**
Overall Score of A+V+T	**0.433**	0.422
Average acc.	**0.462**	0.447

TABLE 6. Comparative evaluation of LSTM models (1) without pretraining and (2) pretrained with fine-tuning. Pretraining was performed using the hiring recommendation label. PT indicates pretraining, and PT/FT indicates pretraining and fine-tuning.

Score	Modality	Without PT	PT/FT
Overall Score	Audio	0.350	**0.478**
	Video	0.443	**0.449**
	Text	0.376	**0.481**
	A+V	0.381	**0.527**
	A+T	0.374	**0.416**
	V+T	0.395	**0.426**
	A+V+T	0.367	**0.433**
Modality Score	Audio	0.562	**0.587**
	Video	0.390	**0.391**
	Text	0.527	0.436

pretrained model. In the former, the pretrained model is used as a weight initializer, and the parameters are trained from scratch using the target domain dataset. In the latter (“freeze” strategy), the weights of the pretrained model are kept frozen except in the last few layers, which are retrained using the target data. In our work, we focus only on fine-tuning the last layer, which is commonly known as the feature extractor.

Table 7 shows a comparative evaluation of the LSTM model using different combinations of modality features and the two network adaptation strategies. All performances are reported using the average balanced accuracy. As seen previously, there is a clear trend that indicates that fine-tuning on all layers achieves consistently better performance than fine-tuning on simply the last layer, except for the multimodal (A+V+T) approach to modeling the overall score, where fine-tuning using the “freeze” strategy gains a balanced accuracy score of 0.059 over fine-tuning all layers. The same outlier can be seen when using only the video modality for modeling modality scores, with a gain of 0.034 for PT/Freeze over PT/FT.

A plausible reason for these performance differences is that by fine-tuning all layers, the gradients are allowed to back-propagate to the pretrained parameters and help capture task-specific adjustments. However, there are still cases in which fine-tuning all layers leads to a degradation in performance, as noted above, possibly due to catastrophic forgetting. Furthermore, deeper analysis is needed to estimate when and what to transfer in TL, with the goal of finding a good initialization that facilitates task learning in the target domain [10].

2) IMPACT OF SOURCE DOMAIN TASKS

Similarity in both the domain and task between the source and target is an important consideration in sequential TL settings [10]. The more similar the tasks and domains are to each other, the greater the improvement that can be expected in the TL scenario. In this work, we investigate the impact of our source domain on task learning in the target domain. Specifically, the incongruity in the label space between the source and target domains deserves a deeper look regarding the efficacy of each type of label in the source domain and its impact on the pretraining and fine-tuning processes in the target domain. More concretely, we pretrained six LSTM models using each of the hiring recommendation and personality label classes to see how they affected the ability of the model to evaluate the overall score of oral presentations.

We present our results in Table 8, where we compare bimodal and multimodal approaches to modeling the overall score using different variants of the source domain label class during pretraining. Additionally, we note the performance of the pretrained LSTM model over a version without pretraining for each modality combination. The balanced accuracy scores indicate yet another clear trend, that pretraining on any of the six label classes in the source domain yields an improvement over the corresponding version without pretraining, regardless of the modality combination.

In the audio+video bimodal model, the model pretrained on conscientiousness achieved the highest balanced accuracy score of 0.561 among all the pretrained models. In multimodal scenarios, the model pretrained on extraversion achieved the highest score of 0.539 among all models. The model pretrained on hiring recommendation showed an improvement of 0.146, followed by the normal supervised model without TL. Meanwhile, the models pretrained on agreeableness, emotional stability, extraversion, and openness showed improvements in the score of 0.052, 0.077, 0.034, and 0.107, respectively. Since the model results changed depending on the source domain labels, we conclude that the stability of the models depends on the pretrained domain and labels. Therefore, choosing related and similar domains or tasks is important for further downstream tasks. In total, in both models (audio+visual and multimodal), the second-best accuracy was obtained by the model pretrained on hiring recommendation, so using the source domain task that estimated the hiring recommendation yielded stable classification performance in this experiment.

C. RESULTS OF THE LARGE-SCALE MODEL (ANSWER TO RQ3)

The results in Table 9 provide a comparative evaluation between our proposed TL model and the model using the pretrained CNN/language model as a feature extractor. The proposed TL methods used the word2vec features for linguistic features, COVARAP features for audio features, and OpenFace features for visual features. Table 9 compares the average balanced accuracy between the proposed TL models trained on the Hiring Recommendation label using the

TABLE 7. Comparative evaluation of LSTM models that were fine-tuned using two strategies: (1) fine-tuning all layers (PT/FT) and (2) fine-tuning on only the last layer (PT/Freeze). Pretraining was performed using the hiring recommendation label.

Score	Modality	PT/FT	PT/Freeze
Overall Score	Audio	**0.478**	0.458
	Video	**0.449**	0.437
	Text	**0.481**	0.463
	A+V	**0.527**	0.364
	A+T	**0.416**	0.395
	V+T	**0.426**	0.389
	A+V+T	0.433	**0.492**
Modality Score	Audio	**0.587**	0.518
	Video	0.391	**0.425**
	Text	**0.436**	0.359

TABLE 8. Comparative evaluation of LSTM models that were fine-tuned using different source domain labels.

Pretraining Conditions	Audio+Visual	Mutimodal
without PT	0.381	0.367
PT on Hiring recommendation	**0.527**	**0.433**
PT on Agreeableness	**0.443**	**0.409**
PT on Conscientiousness	**0.561**	**0.393**
PT on Emotional stability	**0.458**	**0.377**
PT on Extraversion	**0.415**	**0.539**
PT on Openness	**0.488**	**0.390**

TABLE 9. Comparison results with the large-scale feature extractor and intermediate features. Pretraining was performed using the hiring recommendation label for intermediate features. "St-LSTM" denotes Stacked-LSTM.

Modality	Features	Overall Score		Modality Score	
		LSTM	St-LSTM	LSTM	St-LSTM
Audio	YAMNet	0.466	0.443	0.527	0.517
	VGGish	0.400	0.400	0.433	0.433
	proposed TL	**0.478**	**0.532**	**0.587**	**0.535**
Visual	VGG	0.400	0.400	0.383	**0.383**
	MobileNet	0.382	0.333	**0.392**	0.358
	proposed TL	**0.449**	**0.424**	0.391	0.294
Text	BERT	0.377	**0.495**	0.345	0.473
	proposed TL	**0.481**	0.485	**0.436**	**0.538**

TABLE 10. Results of a paired t-test ("(HR)" denotes that the model is pretrained on the hiring recommendation label. "(CO)" denotes that the model is pretrained on the conscientiousness label."(EX)" denotes that the model is pretrained on the extraversion label).

Overall Score	Acc-Improv	t-value	p-value	Hypothesis
Audio (HR)	0.128	1.936	0.085	Accept
Text (HR)	0.105	2.718	0.023	Reject
A+V (HR)	0.146	2.360	0.043	Reject
A+V (CO)	0.186	3.564	0.006	Reject
Multimodal (EX)	0.172	2.497	0.034	Reject

intermediate features and the model trained with the features extracted from the pretrained CNN/language model for each modality. The proposed TL acoustic model outperformed the pretrained CNN model on both scores. For the visual models, the overall score of the proposed TL model performed better, but in terms of the modality score, the pretrained CNN models were better. Last, in the text modality, the proposed TL model trained on LSTM outperformed the pretrained model, which also used the LSTM layer, but when using Stacked-LSTM, the results were reversed. However, in the modality score of the text modality, the proposed TL model clearly outperformed the BERT model. In that case, we can conclude that our proposed TL model performed better in most of the cases than the models using the pretrained CNN/language model. Therefore, using different but similar data to improve the accuracy assessment can be said to be feasible, but we need deeper and more detailed analysis to reach certainty in that case.

VII. DISCUSSION

A. VALIDATION OF THE IMPROVEMENT BY TL

In this section, we confirm that our proposed TL models do not improve the results merely by chance by using statistical hypothesis testing. The hypothesis testing is based on statistical significance, which shows that the observed data are strong evidence against the presumed hypothesis by rejecting the null hypothesis. For hypothesis testing, we use Student's paired t-test to determine the statistical significance of our model. We perform the test on the cases that showed the greatest improvement in accuracy when we adopted our TL techniques. For a given pair of models (before and after applying TL), we determine whether there is a significant difference between the mean balanced accuracy of the same 10 folds used in the cross-validation in the paper. The null hypothesis (that the mean balanced accuracy between a pair of models is the same) is rejected if $p<0.05$. Furthermore, all experiments are conducted with the same random seed to ensure reproducibility. Specifically, we conduct the paired t-test for the modeling scenarios in Table 10. The results show that there are significant differences between the models before and after applying TL for 4 out of 5 modeling scenarios. From the results, we can conclude that our models are statistically significant.

B. TARGET DOMAIN DATA SIZE

To investigate the performance of the proposed transfer learning method with further limited training data samples akin to few-shot learning scenarios, we design experiments to verify the robustness of our TL models under a size constraint. To limit the amount of available training data, we choose 20%, 50%, and 70% of the samples from among all data samples of the target domain. After that, we train the model with a subset of the data and observe the performance on a held-out test split that is unchanged. We limit the training to each training fold of cross-validation, but the test fold of the data remains unchanged. We perform these experiments using a full multimodal model (A+V+T) and LSTM pretrained on a hiring recommendation model. Table 11 shows the

TABLE 11. Investigation of the effect of the target data size on the classification accuracy of LSTM models trained with data subsets in the target domain. Note that the pretrained source model used here is trained using the hiring recommendation label.

	without PT	PT/FT			
Target labels	All	20%	50%	70%	All
A+V+T	0.367	0.435	0.438	**0.473**	0.433

classification accuracy of the LSTM models trained with data subsets of the target domain.

From this table, the degradation of the accuracy of the proposed TL model is alleviated even if the training data size degrades to 20%; the accuracy with 20% of the training data is 0.435, and that of the non-TL method with all training data is 0.367. Overall, it can be seen that using a subset of available data can lead to better performance, so applying TL to a smaller amount of training data also helps in raising performance.

C. AUDIO FEATURE ABLATION

From the experimental results, we found that the most effective modality for the presentation skill assessment prediction task was audio. We investigate the contribution of each type of audio feature to the assessment task. For this purpose, we divide the acoustic features included in COVAREP into three groups:

1) Features related to prosodic information
2) Features related to spectral information
3) Features related to voice quality.

The analysis is performed with an ablation test by removing one feature group from the features (prosodic, spectral, and voice quality). We compare the model lacking the removed feature group with the baseline model trained with all features (All). The details of the audio features are shown in Table 3. Table 12 describes the analyses of the contribution of the feature group to two prediction tasks (Overall and Modality score). 'Excluded' means removed from the features and 'Diff.' means the difference in accuracy ('Acc.') between the current prediction and that obtained using all audio features.

For the prediction of the overall score, these results showed that the voice quality (VQ) is the most effective (the accuracy degraded by 0.004 when excluding VQ). These results indicate that VQ is an important feature for predicting the overall score determined by the coders when using multimodal information from videos.

For the prediction of the modality score, there are no irrelevant feature groups, so the results show that all types of features contributed to the prediction of the modality score. In particular, prosodic and spectral features are the most relevant features for predicting the score (the accuracy degraded by 0.104 when they were excluded). The modality score is judged by observing only speech data, so all audio features that have different aspects capturing speech characteristics are considered to be key descriptors in predicting the score.

TABLE 12. Contribution of each feature group of the audio feature set using our proposed method, determined by an ablation test. The pretrained source model in our method is trained using the hiring recommendation label. 'Excluded' means removed from the features, and 'Diff.' means the difference in accuracy ('Acc.') between the current prediction and that obtained using all audio features.

Audio Features	Overall Score		Modality Score	
All (Prosodic+VQ+Spectral)	0.431		0.577	
Excluded	**Acc.**	**Diff.**	**Acc.**	**Diff.**
VQ	0.427	+0.004	0.540	+0.037
Spectral	**0.541**	-0.110	0.57	+0.020
Prosodic	0.432	-0.001	**0.580**	-0.003
VQ+Spectral	0.436	-0.005	0.546	+0.031
Prosodic+VQ	0.414	+0.017	0.567	+0.010
Prosodic+Spectral	0.475	-0.044	0.473	+0.104

D. FUTURE WORK

In the future, it would be worthwhile to pursue TL scenarios in which the source and domain datasets and objective functions differ significantly as well as to attempt to transfer knowledge from multiple source domains to a single target domain. In the current work, we implement the TL scenarios in the presentation assessment by using only the simple layer sequential model. We do not focus on data augmentation or explore in-depth feature selection and modality combinations for other sources. One conclusion concerns the relatively less effective textual features during TL: due to topic differences between the source and target domains, we believe it is important to ensure that we extract context-independent linguistic features for TL between the domains, rather than using linguistic features that are context-dependent, such as word2vec. Context-independent features from text data, such as discourse coherence and syntactic complexity, can be used as textual features for modeling. Additionally, we are interested in leveraging the growing resources focused on machine learning and AI model interpretation and explainability to apply to our work. For example, we can utilize SHAP [60] (SHapley Additive exPlanations) that is based on game theory to measure feature importance or relevance to predicting a specific instance outcome. Similarly, other libraries such as LIME [61] (Local Interpretable Model-agnostic Explanations) can be used to evaluate the performance of the model against human intuition by examining the features that are used by the model to generate a certain prediction, such as the efforts expended in [6]. Currently, we have not further explored any bias mitigation related to the participant's race, age, or ethnicity when modeling, and would leave those for future work.

VIII. CONCLUSION

In this work, we propose a simple but effective transfer learning framework for improving model performance in evaluating oral presentations in a target domain by utilizing knowledge learned from a different but related task in a source domain using sequential models. Through the comparative evaluation of predictive models that are trained with and without transfer learning, we demonstrate its potential usefulness for a target domain in which (1) there is very little data for training a predictive model from scratch and (2) labels for training such models can be expensive and laborious to obtain. From Table 11, the greatest improvement in the accuracy obtained by TL was 0.106 points (10.6%) (this answers RQ1). Moreover, we investigated several parameters of the TL framework, such as the impact of the similarity of the labels and classes between the source and target domains on the effectiveness of a TL framework. From the results of the investigation of effective TL approaches, fine-tuning all layers effectively improves the accuracy, and using the source domain task that estimates the hiring recommendation yields stable classification performance in this experiment (this answers RQ2). Last, we performed a comparative evaluation between the proposed TL methods and a standard TL method based on a large-scale pretrained model, and the results show that our approach performed better than the standard TL method in most cases. Our proposed model is simple, yet the results suggest that there is promise in implementing TL between two different domains in communication assessment tasks (this answers RQ3).

REFERENCES

[1] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer, ''Multimodal public speaking performance assessment,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 43–50.

[2] V. Ramanarayanan, C. W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft, ''Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 23–30.

[3] O. Hargie, *The Handbook of Communication Skills*, 3rd ed. Evanston, IL, USA: Routledge, 2006.

[4] 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 *Proc. Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 200–203.

[5] L. Hemamou, G. Felhi, V. Vandenbussche, J.-C. Martin, and C. Clavel, ''HireNet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews,'' in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, 2019, pp. 573–581.

[6] C. W. Leong, K. Roohr, V. Ramanarayanan, M. P. Martin-Raugh, H. Kell, R. Ubale, Y. Qian, Z. Mladineo, and L. McCulla, ''To trust, or not to trust? A study of human bias in automated video interview assessments,'' in *Proc. ICCV Workshop Interpreting Explaining Visual Artif. Intell. Models*, 2019, pp.1–6

[7] Y. Yagi, S. Okada, S. Shiobara, and S. Sugimura, ''Predicting multimodal presentation skills based on instance weighting domain adaptation,'' *J. Multimodal User Interfaces*, vol. 16, no. 1, pp. 1–16, Mar. 2022.

[8] G. Murray and C. Oertel, ''Predicting group performance in task-based interaction,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 14–20.

[9] S. J. Pan and Q. Yang, ''A survey on transfer learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[10] S. Ruder, ''Neural transfer learning for natural language processing,'' Ph.D. thesis, School Eng. Inform., Univ. Galway, Galway, Ireland, 2019.

[11] L. Chen, R. Zhao, C. W. Leong, B. Lehman, G. Feng, and M. E. Hoque, ''Automated video interview judgment on a large-sized corpus collected online,'' in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 504–509.

[12] H. Lepp, C. W. Leong, K. Roohr, M. Martin-Raugh, and V. Ramanarayanan, ''Effect of modality on human and machine scoring of presentation videos,'' in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 630–634.

[13] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, ''Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach,'' in *Proc. Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 50–57.

[14] L. S. Nguyen, D. Frauendorfer, M. S. Mast, and D. Gatica-Perez, ''Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior,'' *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1018–1031, Jun. 2014.

[15] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, ''A nonverbal behavior approach to identify emergent leaders in small groups,'' *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 816–832, Jun. 2012.

[16] S. Okada, Y. Ohtake, Y. I. Nakano, Y. Hayashi, H.-H. Huang, Y. Takase, and K. Nitta, ''Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets,'' in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 169–176.

[17] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, ''MACH: My automated conversation coach,'' in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2013, pp. 697–706.

[18] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, and S. Nakamura, ''Automated social skills trainer,'' in *Proc. Int. Conf. Intell. User Interfaces*, Mar. 2015, pp. 17–27.

[19] H. Trinh, R. Asadi, D. Edge, and T. Bickmore, ''RoboCOP: A robotic coach for oral presentations,'' *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 1–24, Jun. 2017.

[20] A. Rosenberg and J. Hirschberg, ''Acoustic/prosodic and lexical correlates of charismatic speech,'' in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Sep. 2005, pp. 513–516.

[21] S. Scherer, N. Weibel, L.-P. Morency, and S. Oviatt, ''Multimodal prediction of expertise and leadership in learning groups,'' in *Proc. Int. Workshop Multimodal Learn. Anal.*, Oct. 2012, pp. 1–8.

[22] M. Chollet and S. Scherer, ''Assessing public speaking ability from thin slices of behavior,'' in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 310–316.

[23] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer, ''Exploring feedback strategies to improve public speaking: An interactive virtual audience framework,'' in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2015, pp. 1143–1154.

[24] M. Chollet, H. Prendinger, and S. Scherer, ''Native vs. non-native language fluency implications on multimodal interaction for interpersonal skills training,'' in *Proc. ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 386–393.

[25] M. Chollet, T. Massachi, and S. Scherer, ''Racing heart and sweaty palms,'' in *Proc. ACM Int. Conf. Intell. Virtual Agents (IVA)*, 2017, pp. 83–86.

[26] E. Kimani, P. Murali, A. Shamekhi, D. Parmar, S. Munikoti, and T. Bickmore, ''Multimodal assessment of oral presentations using HMMs,'' in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 650–654.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ''ImageNet large scale visual recognition challenge,'' *J. comput. Vis.*, vol. 115, pp. 211–252, 2015.

[28] X. Qiu et al., ''Pre-trained models for natural language processing: A survey,'' *Sci. China Technol. Sci.*, vol. 63, pp. 1872–18972, 2020, doi: 10.1007/s11431-020-1647-3.

[29] O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn, ''Cross-lingual word embeddings for low-resource language modeling,'' in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 937–947.

[30] Y. Wang, Y. Sun, Z. Ma, L. Gao, Y. Xu, and T. Sun, ''Application of pre-training models in named entity recognition,'' in *Proc. 12th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, vol. 1, Aug. 2020, pp. 23–26.

[31] J. Howard and S. Ruder, ''Fine-tuned language models for text classification,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339, doi: 10.18653/v1/P18-1031.

[32] K. Feng and T. Chaspari, ''A review of generalizable transfer learning in automatic emotion recognition,'' *Frontiers Comput. Sci.*, vol. 2, p. 9, Feb. 2020.

[33] Q. Li and T. Chaspari, ''Exploring transfer learning between scripted and spontaneous speech for emotion recognition,'' in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 435–439.

[34] H. Li, Y. Kim, C.-H. Kuo, and S. Narayanan, ''Acted vs. improvised: Domain adaptation for elicitation approaches in audio-visual emotion recognition,'' in *Proc. Interspeech*, 2021, pp. 3395–3399, doi: 10.21437/Interspeech.2021-666.

[35] X. Wang, C. Tang, X. Zhao, X. Li, Z. Jin, D. Zheng, and T. Zhao, ''Transfer learning methods for spoken language understanding,'' in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 510–515.

[36] Y. Yin, B. Huang, Y. Wu, and M. Soleymani, "Speaker-invariant adversarial domain adaptation for emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 481–490.

[37] G. Boateng, L. Sels, P. Kuppens, P. Hilpert, and T. Kowatsch, "Speech emotion recognition among couples using the peak-end rule and transfer learning," in *Proc. Companion Publication Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 17–21.

[38] G. Boateng and T. Kowatsch, "Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 12–16.

[39] A. Rassadin, A. Gruzdev, and A. Savchenko, "Group-level emotion recognition using transfer learning from face identification," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Glasgow, U.K. New York, NY, USA: Association for Computing Machinery, 2017, pp. 544–548, doi: 10.1145/3136755.3143007.

[40] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 443–449.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: Association for Computational Linguistics, doi: 10.18653/v1/N19-1423.

[42] D. Hazarika, S. Poria, R. Zimmermann, and R. Mihalcea, "Conversational transfer learning for emotion recognition," *Inf. Fusion*, vol. 65, pp. 1–12, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253520303018, doi: 10.1016/j.inffus.2020.06.005.

[43] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly fine-tuning 'BERT-like' self supervised models to improve multimodal speech emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2020, pp. 3755–3759.

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

[45] S. S. Y. Tun, S. Okada, H.-H. Huang, and C. W. Leong, "Analysis of modality-based presentation skills using sequential models," in *Social Computing and Social Media: Experience Design and Social Network Analysis*. Cham, Switzerland: Springer, 2021, pp. 358–369.

[46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, 2013, pp. 1–12.

[47] S. Bird and E. Loper, "NLTK: The natural language toolkit," in *Proc. ACL Interact. Poster Demonstration Sessions*, Barcelona, Spain, 2004, pp. 214–217.

[48] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*, Valletta, Malta, 2010, pp. 46–50.

[49] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 960–964.

[50] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.

[51] K. Weiss, T. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, 2016.

[52] P. Diederik Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–15.

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, Jun. 2014.

[54] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.

[55] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

[56] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, pp. 1–10, 2016.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp.1–14.

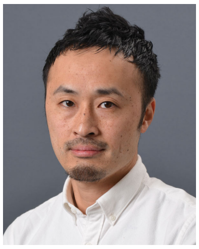
[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[59] G. Andrew Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, pp. 1–92017.

[60] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 1–10.

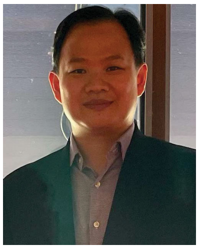
[61] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

SU SHWE YI TUN received the B.S. degree from the University of Information Technology, Myanmar, in 2019, and the M.S. degree from the Japan Advanced Institute of Science and Technology (JAIST), in 2021. Her current research interests include social signal processing and multimodal interaction.



SHOGO OKADA (Member, IEEE) received the Ph.D. degree from the Tokyo Institute of Technology, Japan, in 2008. In 2008, he joined as an Assistant Professor with Kyoto University. In 2011, he joined as an Assistant Professor with the Tokyo Institute of Technology. He joined the IDIAP Research Institute, Switzerland, as a Visiting Faculty Member, in 2014. He currently directs the Social Signal and Interaction Group, Japan Advanced Institute of Science and Technology (JAIST), Japan, where he is also an Associate Professor. His current research interests include social signal processing, human dynamics, multimodal interaction, and machine learning. He is a member of ACM, JSAI, and IEICE.

HUNG-HSUAN HUANG received the Ph.D. degree in informatics. He is currently the Director of the Laboratory for Interactive Intelligent Systems and a Full Professor with the Faculty of Informatics, The University of Fukuchiyama, Fukuchiyama, Kyoto, Japan. He will mentor papers in the areas, such as interaction design with robots, techniques in the perception of robots, generation of robots' behaviors, multimodal dialogue systems, and evaluation of the interaction. His current research interests include applied artificial intelligence, human–computer interaction, communication science fields, multimodal interaction with virtual agents, and communication robots.



CHEE WEE (BEN) LEONG received the Ph.D. degree in computer science and engineering. He is currently a Managing Principal Research Engineer with the AI Laboratories, Educational Testing Service (ETS), USA, where he directs, plans, and implements prototyping of multimodal AI products and services in the educational domain and learning space. His current research interests include multimodal modeling of affective states, evaluation and summarization of noncognitive skills, and multimodal data fusion for behavioral trait prediction.