A Social Condition-Enhanced Network for Recognizing Power Distance using Expressive Prosody and Intrinsic Brain Connectivity

Fu-Sheng Tsai, *Student Member, IEEE,* Wei-Wen Chang, Chi-Chun Lee, *Senior Member, IEEE*

***Abstract*—Culture is the social norm that often dictates a person's thoughts, decision-making, and social behaviors during interaction at an individual level. In this study, we present a computational framework that automatically assesses an individual culture attribute of power distance (PDI), i.e., the measure to describe one's acceptance of social status, power and authority in organizations through multimodal modeling of a participant's expressive prosodic structures and brain connectivity using a social condition-enhanced network. In specific, we propose a joint learning approach of center-loss embedding network architecture that learns to "centerize" the embedding space given a particular social interaction condition to enhance the PDI discriminability of the representation. Our proposed method achieves 88.5% and 73.1% in binary classification task of recognizing low versus high power distance on prosodic and fMRI modality separately. After performing multimodal fusion, it improves to 96.2% of 2-class recognition rate (7.7% relative improvement). Further analyses reveal that average and standard deviation of speech energy are significantly correlated with power distance index; the right middle cingulate cortex (MCC) of brain region achieves the best recognition accuracy demonstrating its role in processing a person's belief about power distance.**

***Index Terms*—culture dimensions, fMRI, prosody, center-loss embedding, power distance index**

I. Introduction

In human society, *culture* is a complex construct resulting from a composition of primitive attributes such as belief, knowledge, customs, and many more common habits that human possesses in shaping the society as a whole [1]. Anthropologists have long considered that the formation of culture is based on human's unique capability in transmitting and influencing each other's habits and thoughts in a given society through social learning [2]. The culture emerges from ecological social conditions and drives behaviors which affects our social goals, morals and mindsets toward life. It has been shown that under different social settings, the culture guides our action and decision toward life at an unconscious level and shapes up to the diverse organizations [3], [4]. This particular social aspect also underlies each individual's behaviors, social interaction dynamics, modulates the development of personality and attitudes, and affects our decision toward life events. To better describe the effects of culture and its relationship to human behaviors, Hofstede develops the cultural dimension theory that has been widely used as a core framework for understanding cross-cultural communication [5], [6]. Hofstede states that there are six dimensions of cultures, including power distance index (PDI), individualism vs. collectivism (IDV), uncertainty avoidance (UAI), masculinity vs. femininity (MAS), long-term orientation vs. short-term orientation (LTO) and indulgence vs. restraint (IND). Many research has been dedicated in investigating and employing Hofstede's cultural dimensions to understand how these values affect our life, e.g., Taras et al. examine the use of Hofstede's cultural value dimensions for predicting organizational and employee performance outcomes [7]. Among the six dimensions of Hofstede's culture, PDI refers to the cultural construct indicating how the power is distributed and the extent in which social inequalities are accepted and viewed as natural in a society [6], [8]. It can further be used to describe an individual's belief about power, authority and status in organizations [9].

PDI often varies from culture to culture, which makes it as a critical component in studying inter-cultural communication. People in society with high PDI tend to conform with the social stratification, which means "everyone has a social place that needs no further justification" [10]. In contrast, people in low PDI society prefer more equal status, interactions and inalienable rights to distribute power equally, where phenomenon of social inequality requires more justification and democratic consent. The effect resulting from differences of PDI is often observable when people engage in interactions under different conditions of social status. For instance, people in low PDI condition exist open door policy which superiors are not only open to inferiors, but subordinates are also more likely to be willing to challenge or suggest to their superiors; on the contrary, subordinates in high PDI society are unlikely to approach and contradict their bosses directly [5].

In the domain of human-centered multimedia processing, culture provides additional behavior modulation manifested in nonverbal and supplementary cues which leads to better communication of human in memory, understanding and perception. Studies have claimed the importance in considerig culture factors [11] or personality [12], [13] since behaviors and signals that human generates depend on personal cultural background and interaction contexts. Furthermore, studies in [14] and [15] indicate culture as well as personality do influence one's perception on video, e.g., different contexts in which participants are induced would result in varying expressions, which should be taken into consideration in developing computational modeling. Researchers additionally have

Wei-Wen Chang is with the Department of International Human Resource Development, National Taiwan Normal University, Taiwan

Fu-Sheng Tsai and Chi-Chun Lee are with the Department of Electrical Engineering, National Tsing Hua University, Taiwan and MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

already examined impacts of human factors such as cultural and personality traits, to understand their effects toward human daily life and interactions. For instance, Guntuku et al. [16] have proposed algorithm in Quality of Experience (QoE) modeling and management with comprehensive investigations on contexts and human factors of QoE system. They have shown that personal and cultural traits are significant impact factors in predicting QoE, especially on the intensity of negative influences. In more recent works, Varini et al. [17] propose a customized egocentric video recommendation system based on 3D Convolutional Neural Network (CNN) approach to embed visual apparent motion and real motion feature representations together. By training with each user's preference, the visual semantic classifier assesses the correlation between extracted key shots in cultural heritage scenario and users' preferences. Consequently, this study also points out the importance of quantitatively modeling the human behavioral data with one's cultural and personal traits.

While many of these researches have demonstrated effectiveness of using Hofstede's culture dimensional framework in understanding human communication, there still exists limitation in quantifying cultural constructs based on conventional self-reported instruments. For example, while Hofstede's culture framework has been used as a world-wide instrument in quantifying cultural constructs specifically relevant in business setting, the large-scale validity of such a quantification scheme is often at doubt. Even the current largest scale studies (e.g., [18]–[20]) conducted in numerous countries draw conclusions relying on few sample data points collected with a limited longitudinal analysis time span. Additionally, recent works in cross-cultural study have started to investigate methods in improving Hofstedes systems of culture measurements, e.g., Taras et al. [21] propose to improve national-wide cultural indices by using longitudinal meta-analysis of Hofstedes culture dimensions in deriving country-wide cultural ranking, which is shown to provide a more accurate and advanced guidance for managers to optimize the performance of their organization in cross cultural conditions. Moreover, most previous works regard cultural power factor as a unidimensional construct, which point out the limitations of using self-report instruments. For example, research in [22] mentions that the main issue in affecting the reliability and validity of using self-report measurement is the social desirability bias, i.e., participants would often respond in a socially acceptable way. For example, individuals may be more likely to response "yes" or "no" regardless of question content that is known as acquiescent or non-acquiescent, respectively. Furthermore, [23] claims that the use of self-report instruments lacks measurement invariance from participants across seven nations and needs further modification for cross-cultural comparison. On the other hand, third person ratings through observation have been widely used due to limitations of self-report instrument. However, study in [24] also demonstrate issues of validity and reliability of using manual (human-based) observational measurements, such as inter-rater, intra-rater, test-retest reliability, internal consistency (i.e., measurements unidimensionality should be checked) and measurement error.

We argue that developing computational modeling technique that is capable of objectively quantifies behaviors and directly models signals of interest in a data-driven approach should be taken along with these existing scales for measuring the complex cultural construct not only to improve the validity and reliability but also to enable longitudinal studies potentially across nations at scale. Hence, aside from the issue of continuous modification on the definition of cultural dimensions, many of these issues in studying cultural constructs for human communication still suffer from a lack of objective methods that hinders systematic large-scale studies and even adoption to real-world applications. Unlike psychological constructs of emotion states, where a tremendous algorithmic development has progressed well over a decade since the term of "affective computing" was first proposed [25]. This large body of computational studies concentrate on developing automated methods in recognizing emotion from measurable human signals using machine learning techniques, e.g., performing emotion recognition using speech [26], [27], facial expressions [28], body gestures [29], [30], neuro-responses or physiological signals [31], [32] and multimodal behaviors [33], [34].

While cultural dimensions impact our daily life and behaviors during interpersonal communications, limited, if any, researches has dedicated to automatic computation of an individual's cultural dimensions from measurable behaviors. Our previous work has demonstrated that PDI measure of an individual could be assessed automatically by modeling the interlocutor's prosodic expressions during a face-to-face conversation when considering its differential manifestations simultaneously across different social conditional settings [35]. Prosodic structure during human spoken dialog not only reflects our behaviors and internal states but implies discriminatory power of attitudinal expressions of social interaction. Our previous work has proposed a first attempt in deriving objective method to automatically recognize PDI dimension through speech prosody. In this work, we continue to extend by proposing a framework in integrating both expressive vocal characteristics and internal brain connectivity to advance the PDI recognition accuracy. In the following, we will list relevant works and summarize the specifics of our contribution in this work.

A. Related Works

Spoken dialog is the most natural form of our real-world interactions and communications in our social life. It is the most important vehicle for humans to transmit social intent to gain access to each other's need, and the expressive behavior manifestations of these social exchanges inevitably are modulated by one's cultural background. While there has not been direct computational works in linking behaviors in spoken interaction to one's cultural background, a number of research works have already pointed out that the expressive aspects of speech prosody can be conceptualized as realization of two different processes during spoken dialogs: the involuntarily-controlled expressions of affects - the so-called emotions, and the intentionally-controlled attitudinal functions of social factor - the so-called attitude [36]. Research in [37] has shown that low power individuals pay more attention to acoustic

fluctuations than high power ones, resulting in higher emotional prosody recognition accuracy which corroborates recent researches [38] in indicating low power distance individuals increase vigilance during the processing of perceptual cues. These past works demonstrate that prosodic variation not only indicates human internal states, e.g., mental states, emotion, mood, but also significantly affected by social-cultural factors between dialog partners [39]–[41]. Furthermore, West et al. [42] have used linguistic-based measurement of cultural distance, which is based on languages genetic classification; they report that these measurements are representative among members of society and easy to operationalize for all languages, and are additionally significantly correlated to the managerial values. Very few, if any, studies have investigated paralinguistic aspects in a face-to-face conversational setting. Except for study in [43], it has been found there exists phonetic and paralinguistic differences for situation in talking to people with polite vs. impolite social status. These suggest that both expressive linguistic and paralinguistic measurements carry information about one's own cultural attribute of social distance constructs.

Specifically, the relationship between attitudinal expressions and social-cultural backgrounds has been quite extensive studied. For example, Mixdorff et al. and Barbulescu et al. [39], [44] propose to use macro-prosodic parameters to distinguish different types of social attitudes for each specified culture. At the same time, since the variation of cultural background affects realization of attitudinal expressions in social settings, Shochi et al. investigate the effect of prosodic parameters for inter-cultural (English, Japan and French) perception of affect [45]. Aside from examining expressive vocal characteristics, understanding the internal brain connectivity as a function of social hierarchies, i.e., perception of social status when interacting with dialog partners of different status, has also been separately investigated in the field of neuroscience with the availability of functional magnetic resonance imaging (fMRI) technique. For example, Koski et al. [46] explore the nature of social hierarchies and characteristics associated with social status for both human and nonhuman; their findings indicate that brain activity could be used to rapidly recognize the social status. In addition, Liew et al. [47] propose to examine how cultural differences affect an individual's implicit self-processing in different social conditions; their experimental results demonstrate that several brain regions are activated during observation of signals related to social dominance, indicating the differences in social attitude are personally rewarding.

1) Our Contributions: On the expressive side, while research has shown that speech prosody plays an important role in distinguishing social attitudes, limited works have considered culture-prosody relationship at an individual level. At the same time, while most of the past neuro-scientific frameworks point out the strong relationship between social hierarchies and brain activity, there has not been any principled modeling techniques attempting to automatically recognize an individual's culture dimension from these collected brain images.

In our previous work [35], we propose a social condition-enhanced prosodic network (SC-ePN) that models an individual's expressive prosodic structure by simultaneously considering its manifestation over three variants of social conditions. SC-ePN obtains an initial promising accuracy in classifying high versus low PDI index of an individual. In this work, we extend beyond our previous work specifically with the following additional contributions:

1) Multimodal recognition: develop a multimodal, i.e., expressive prosody characteristics with internal brain connectivity, within a social condition-enhanced network, i.e., through inclusion of a center loss constraint integrated with respect to the exposed social conditions, to obtain improved classification rates between high versus low PDI cultural dimension.
2) Analyses: investigate the important prosodic variables and brain regions in automatically inferring the PDI measures.
3) Visualization: demonstrate the effect of our use of center loss in constraining the learning of prosodic and brain connectome networks embedding representations.

In short, the aim of this work is to automatically assess personal culture value of PDI to further understand quantitatively how an individual's belief on power distribution in society would be manifested in his/her attitudinal prosodic structure during social conversation and also in the functional activation within the brain when being exposed to different social encounters. There already exists a number of computational works in assessing different complex human internal constructs, e.g., affective state recognition [48], personality assessment [49], and detection of social disorder [50], but very limited, if any, work has worked on cultural constructs, especially not in the multimodal context. By introducing the use of our proposed deep social-condition enhanced network, which not only models both modalities simultaneously but also integrates information of such signal across different conditions (where it is known that the culture value would likely to impact these measurable human data), we present one of the first works that demonstrate a high recognition accuracy (96.2%) and systematically analyze the difference on each of these modalities for subjects with high versus low PDI value.

The rest of the paper is organized as follows: Section 2 describes about our multimodal database, collection methodology, and power distance measurement. Section 3 describes about research methodologies including multimodal feature representation and our social condition-enhanced network. Section 4 shows our experimental setups, results, analyses and discussions. Section 5 concludes with future works.

II. Database

A. The Multimodal Social-Distance Database

In this section, we will describe the dataset used in this work, which includes both audio and fMRI data collection. We recruit a total of 26 right-handed participants with normal or corrected to normal vision for conducting both audio and fMRI. The collection of 26 subjects has been demonstrated to be practically sufficient for study in using fMRI-based experimentation [51], [52]. All of the subjects are native

Mandarin-speaking participants, who are students attending graduate or undergraduate school (11 females, 15 males, 20-35 years old, mean = 23.23 years old, SD = 3.32). No person with dysphasia, neurological or psychological disorder is included in this study. Each of them is required to finish both audio and fMRI experiments. Each subject is instructed to fill in basic information first. In the audio experiment, the protocol involves asking the target subject to imagine interacting in face-to-face conversation with two persons of different power status, i.e., one as high power status comparing with him/her (e.g., professor or teacher in school) and another as a slightly higher or similar status (e.g., classmate or friend). After conducting experiments of audio recording, the subject is instructed to become familiar with button-pressing device during fMRI scanning. Then each subject performs block design fMRI scanning experiment to record their brain functional activity.

1) Audio Data Collection: We design the audio experimental protocol to investigate the relationship between expressive acoustic cues and an individual's cultural attribute. The experiment protocols ask the participant to imagine a person of two different power status to do something for them. The two social roles are designed as the subjects respected professor/teacher (higher power status) and a senior classmate/TA (similar power status). The former is the role designed such that a participant would respect or even feel stressed and nervous when thinking about meeting with him/her, and the latter is a relaxing situation designed to resemble when a participant imagine interacting with his/her classmate or friend. Then our experimenter would play the role of this person to engage in dialog with the subject. There are 7 questions covering different topics of interaction, and 2 social settings (1 higher power status, 1 similar) for each subject. The following is a list of the 7 conversation topics used in the database to carry out the conversation:

1) There are two free meal tickets, how would you invite him/her to join you?
2) You don't know how to do your homework, how would you ask him/her for help?
3) There is a job interview next week and you want some advice, how would you say to him/her?
4) After you have a fight with someone, you want to seek advice in handling the aftermath, how would you request for help?
5) Your family run into financial difficulties and you are considering about quitting school to find a full-time job, how would you seek advice from him/her?
6) You fail the class with 3 points short preventing you from obtaining the final graduation credits, how would you ask for more points from him/her?
7) Your graduation exhibition is scheduled to take place next week, how would you invite him/her to attend?

The 26 participants result in the audio dataset with a total of 364 (26*2*7) dialogs recorded in this experiment.

2) fMRI Data Collection and Pre-processing: Another goal of the study is to investigate the discriminative power of functional neural connectivity associated with processing social distances when experiencing contexts of different power status. We perform a block design in which targets power status

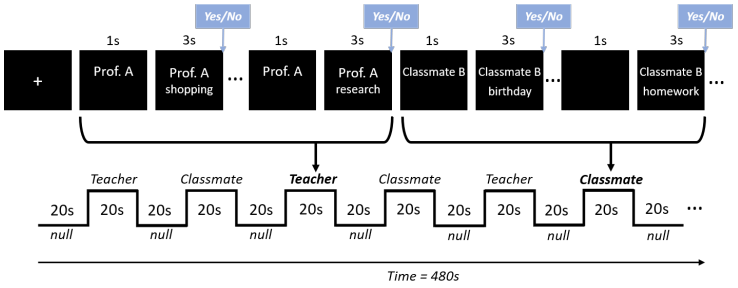


Fig. 1: *It shows the paradigm scheme of fMRI block design experiment. Both the block and null duration are 20 seconds. There are a total of 4 blocks for each target social setting (classmate, teacher) in each run with randomized order. One run contains 8 event blocks and 8 nulls in total. For each block accompanies with 1 null to return BOLD signal to baseline, participant first watches the social setting for 1 second and then responds with their choice of Yes/No of the activity with target person during the 3 second stimuli. Therefore, there are total (8(blocks)+8(null))*20(seconds) = 320 seconds in this fMRI block design experiment.*

person are displayed on the screen. Participants are instructed to watch the visual stimuli and then decide whether he/she would agree to do activities with the targeted status person. The status varies from the acquaintance met often in class or a professor with authoritarian image in school. There are forty social activities of daily experience, for example, party, shopping or gathering invitation, and dissertation discussion. That is, each block shows the target status person paired with different activities, and each participant is equipped with button-pressing device (Yes/ No) hold in their right hand to give their answer. The subjects are not informed about the detailed contents of the experiments a-priori and have a brief practice session for the task outside the MRI scanning room to ensure they could understand the protocol.

MRI scanning is conducted on a 3T scanner (Prisma, Siemens, Erlangen, Germany). T1 sequence (magnetization prepared rapid acquisition gradient echo, MPRAGE) with field-of-view of $240 \times 240\ mm$, 256×256 matrix size and 192 slices are utilized for collecting high resolution structural images. Functional images (task-based images and resting state images) are acquired with gradient-echo EPI sequence to capture the 168 whole-brain images in two runs with TR= 2000ms. We perform all necessary pre-processing steps on the collected MRI data using SPM12 including correction for motion, slice-timing acquisition differences (references to the first slice), coregister to T1-weighted MR images, normalize to standard anatomical template space and spatially smoothing. Additionally, we perform interpolation to generate image sample at 1 second time step to capture the continuous variation within each stimulus.

B. Power Distance Index (PDI) Measurement

Our goal of this work is to automatically assess an individual's power distance index using multimodal data. Power distance is first developed by Hofstede describing the extent

where power inequalities is viewed as natural in society. Furthermore, Sharma proposes that power distance consists of two dimensions: power (POW) and social inequalities (IEQ). POW indicates how individuals are related to authority while IEQ describes ones' hierarchical or egalitarian orientation. In this work, we use POW scale as our numerical measure of participant's perception about authority and social interaction in power relations to measure his/her PDI. POW involves each subject to answer the 7-point Likert scale (ranging from 1 = strongly disagree to 7 = strongly agree) questionnaire used to assess a participant's belief in power of social condition as following:

1) I easily conform to the wishes of someone in the higher position than mine.
2) It is difficult for me to refuse a request if someone senior asks me.
3) I tend to follow orders without asking any questions.
4) I find it hard to disagree with authority figures.

After adding up scores from above questions, we binarize the POW scale in our dataset, where 13 of them are considered as having high PDI (higher than mean value), and rest of 13 are assigned to low PDI (lower than mean value).

III. Research Methodology

Our complete automatic PDI assessment framework is shown as Figure 2. It consists of three components: 1) prosodic and fMRI feature extraction, 2) center-loss social condition-enhanced network, 3) power distance classification with multimodal fusion. In the following sections, we will describe the detailed approach of the three components mentioned above.

A. Acoustic Dynamic Prosodic Features

In this work, we extract the following 13 dynamic prosodic frame-level features proposed by [53] with Praat package [54] from participant's speech during each interaction scenario mentioned above.

- 1 Duration of the voice segment
- 6 Coefficients of 5-degree polynomial function to model pitch contour
- 6 Coefficients of 5-degree polynomial function to model energy contour

We extract pitch and energy at 10 ms intervals and break the contours into pseudo-syllabic segments, and approximate segments of pitch and energy contours by using Legendre polynomial expansions. The arguments for calculating pitch in Praat are shown in Table I. In this section we will describe two steps for dynamic prosodic feature extraction: 1) segmentation, and 2) contour approximation.

1) Segmentation: After we obtained pitch and energy contour from Praat, we segment long pitch contours into syllable-like regions by detecting the valley points of the energy contour [55]. These valley points generally serve as segment boundaries and we impose minimum duration constraint of 60 ms to avoid making a segment too short, which enables us to calculate Legendre polynomial expansions with six terms.

TABLE I: *Parameters of pitch extraction setting in Praat*

Args.	Value
Analysis window length (ms)	30
Time step (ms)	10
Pitch floor (Hz)	75
Pitch ceiling (Hz)	350
Silence threshold	0.03
Voicing threshold	0.6
Voice/Unvoiced cost	0.14
Octave cost	0.01
Octave-jump cost	0.6
Number of candidates (max)	5

2) Contour Approximation: For each segmented contour $f(t)$, we carry out an approximation of pitch and energy contour by taking M-th order Legendre polynomial expansion, which is approximated as

$$f(t) = \sum_{i=0}^{M} c_i P_i(t) \tag{1}$$

where $P_i(t)$ is i-th order Legendre polynomial, c_i is i-th order coefficient, and we set $M = 5$ in this work. Notice that each coefficient represents a particular aspect of the contour. c_0 stands for mean of the contour, c_1 stands for slope of the contour, c_2 stands for culvature of the contour, and c_3, c_4, c_5 model the fine detail. For each segment, we use these six coefficients to be the contour representation. Combing coefficients of pitch and energy, with the segment duration, we obtain a 13 dimensional feature vector. And we further perform context expansion per frame to obtain a total of 39 dynamic prosodic features. These features are then z-normalized with respect to each speaker.

B. Neural Connectivity Graph Embedding

Recent works in cognitive neuroscience have repeatedly demonstrated the importance in quantifying dynamical patterns of inter-regional brain connectivity from both resting state and task-evoked fMRI to conduct neuro-scientific studies [56], [57]. A recent deep representation learning method in characterizing brain functional connectivity is through graph embedding approach [58], which has shown its superior modeling capacity in representing brain connectivity. In this work, we also propose to use a neural connectivity graph embedding approach that follows closely this particular approach.

Specifically, we make use of anatomical automatic labeling (AAL) to split the brain into 90 region of interest (ROI) [59] and calculate the mean of each region to be a ROI-level descriptor. For each scanning session, we use a sliding window approach with window length of 5 seconds and step length of 3 seconds. A 90 x 90 connectivity matrix is then calculated using Pearson correlation coefficient as a measure of inter-region connectivity which refers with the following equation:

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} \times C_{jj}}} \tag{2}$$

where R and C are correlation coefficient and covariance, respectively. And we consider only positive value of correlation and negative value is set to zero since spatial correlations

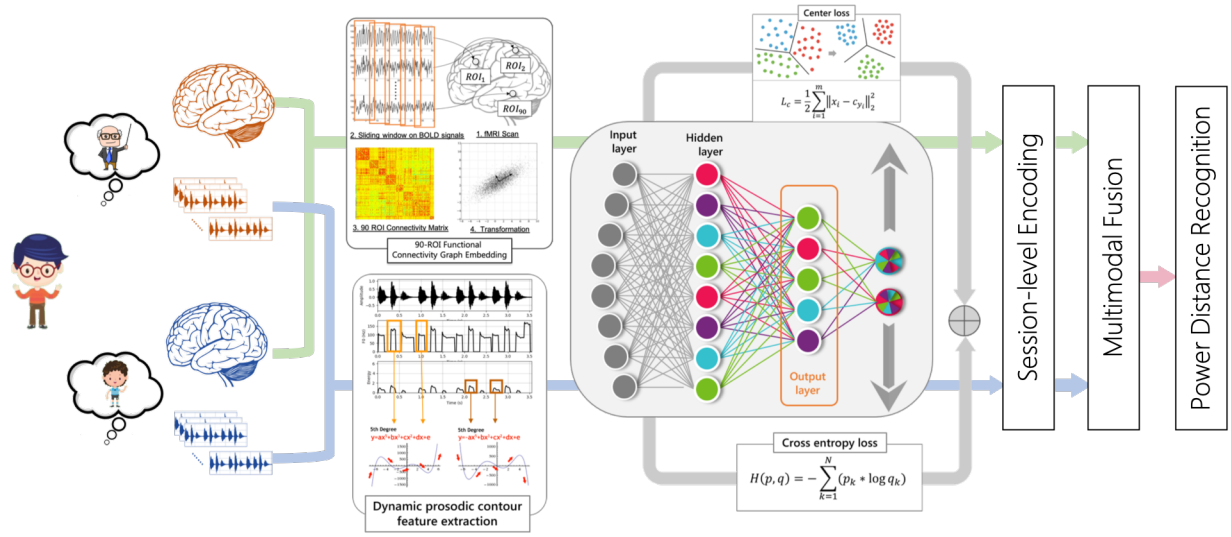


Fig. 2: *It shows the complete architecture of our multimodal social condition-enhanced network (SC-eN) for power distance recognition: ROI-based functional connectivity graph embedding, dynamic modeling of prosodic pitch and energy contour, training networks by jointly optimizing setting-wise center-loss with cross entropy criteria, performing recognition using functional encoding of network output with support vector machine.*

are viewed as edge weights. Finally, we obtain R^G high-dimensional graph embedding descriptors, where G is m*(m-1)/2 and m=90. In order to construct into smaller number of uncorrelated variables of features, we conduct principle component analysis (PCA) by setting the components to be 50 to obtain a reduced dimension of neural connectivity representation.

C. Social Condition-Enhanced Network

Since culture construct, i.e., PDI as measured by POW specifically in this work, either modulates the prosodic expressions or results neural connectivity in a non-linear yet subtle manner, in order to enhance the discriminatory representation power of these modalities, we leverage the natural differences of the manifested signal from the subject's expressed/measured prosodic/fMRI data in different social conditions (as explicitly primed during the experimental setup, i.e., interacting with high and low power status person). Specifically, by introducing the use of center-loss that constrains the modality-specific embedding learning to "centerize" the representation per condition by considering the two social settings simultaneously, we can effectively enlarge the discriminatory information of these embeddings toward power distance classification.

We propose to learn two different social condition-enhanced networks, one for expressive prosodic network (SC-ePN) and one for internal neural connectivity network (SC-eNN), with the use of center-loss constraint. The use of center-loss embedding has recently been applied to various recognition tasks when integrated with neural network embedding; exemplary applications include face recognition [60]–[62], emotion recognition [63], [64], and handwritten Chinese character recognition [65], [66]. In this work, in order to uncover culture value of power distance under different social conditions, we perform joint optimization of both power distance and social conditions, i.e., two optimization targets. The first optimization is the target recognition attribute of power distance index as measured by POW scale, L_{CE}. And the center-loss lists as follow is used to optimize across different social conditions to further constrain the hidden layers:

$$L_c = \frac{1}{2}\sum_{i=1}^{m}\|x_i - c_{y_i}\|_2^2 \tag{3}$$

where m is the number of training samples in a batch. x_i is the i^{th} training sample. y_i is the class (social setting) corresponding to x_i. c_{y_i} denotes as the class center of y_i^{th}. The SC-ePN and SC-eNN both includes three fully-connected layers as shown in Figure 2. The complete loss function in learning condition-enhanced network is a combination of center-loss L_c that centralizes the social setting-specific feature space and L_{CE} learns to classify between high and low power distance.

$$L_{Total} = L_{CE} + \lambda L_c \tag{4}$$

where λ refers to the centering degree and weighting between two losses. We set 0.5 and 0.8 for SC-ePN and SC-eNN, respectively.

D. Multimodal Power Distance Classification

Each of our social condition-enhanced network outputs frame-level feature embedding of each subject's interaction scenario. Due to different length of every session, it results in a varying number of sequences. We additionally apply functional encoding to generate the final feature vector of each participant's session-level representation inputted to the classifier by computing 15 statistical functions. The list of functions includes maximum, minimum, mean, median, standard deviation, 1^{st} percentile, 99^{th} percentile, $99^{th} - 1^{st}$ percentile, skewness, kurtosis, maximum position, minimum position, upper quartile, lower quartile and interquartile range.

After encoding frame-level features, we employ two different multimodal fusion techniques to integrate feature representations from different social settings of different modalities.

TABLE II: It summarizes the Unweighted Average Recall (UAR) obtained in our proposed power distance recognition experiment. P and C denote two different kinds of social condition setting, i.e., professor and classmate. DN indicates representation derived from feed-forward neural network without center-loss, and CN is our proposed social condition-enhanced prosodic network. Re-H/L represents categorical recall.

		Audio (A)		**fMRI (f)**						**Multimodal**		
		DN	CN	DN PCA	CN PCA	DN Graph	CN Graph	DN Graph PCA	CN Graph PCA	(A)CN+ (f)CN PCA	(A)CN+ (f)CN Graph	(A)CN+ (f)CN Graph PCA
P	Re-L	0.462	0.923	0.846	0.538	0.769	0.615	0.615	0.615	0.923	0.769	0.846
	Re-H	0.528	0.846	0.769	0.769	0.385	0.615	0.538	0.462	0.923	0.846	0.923
	UAR	0.5	0.885	0.808	0.654	0.577	0.615	0.577	0.538	0.923	0.808	0.885
C	Re-L	0.692	0.615	0.769	0.385	0.846	0.846	0.692	0.769	0.769	0.692	0.692
	Re-H	0.692	0.615	0.769	0.615	0.462	0.692	0.615	0.846	0.692	0.615	0.846
	UAR	0.692	0.615	0.769	0.5	0.654	0.769	0.654	0.808	0.731	0.654	0.769
$P+C$	Re-L	0.615	0.923	0.769	0.769	0.769	0.846	0.769	0.769	0.769	0.846	0.923
	Re-H	0.615	0.846	0.615	1	0.538	0.538	0.538	0.692	0.923	0.846	1
	UAR	0.615	**0.885**	0.692	**0.885**	0.654	**0.692**	0.654	**0.731**	0.846	0.846	***0.962***

One is based on early-fusion technique, i.e., concatenating audio and neural features into one feature vector after performing univariate feature selection (i.e., ANOVA) on each modality separately. This can be seen as unimodal classification which it obtains the final feature vector before learning and classification steps. Another technique is based on late-fusion that is sometimes called classifier fusion or multimodal classification, i.e., by fusing the decision scores from audio and neural modalities from classifier. The classifier selected in this work is linear-kernel support vector machine that performs the final social power distance recognition for each participant.

IV. Power Distance Recognition Experiment

In this section, we present our recognition results on binary classification between high and low power distance culture dimension. Recognition accuracy is measured in unweighted average recall (UAR) with the evaluation scheme done via leave-one-person-out cross-validation. In further analyses, we investigate the relationship between individual power status and prosodic or neural connectivity measurements.

A. Experimental Setup

The network architectures of acoustic and neural connectivity modality are listed as following table including network layers, batch size, epoch, iteration and weight of center loss. The complete network is trained using Adam (lr = 0.001). We extract the hidden layer of 16-dimensions as participant's acoustic and neural connectivity representation at the frame level. We compare our center-loss embedding network (CN) to a network without center-loss embedding (DN) and list our models from each modality:

- (A) Prosody: Compute 39 dynamic acoustic features and learn a prosodic network with the structures with (CN-Prosody) or without (DN-Prosody) center-loss embedding in deriving the frame-level prosodic representation.
- (f) PCA: Perform principle component analysis on extracted 90-ROI features, and then learn the network with (CN) or without (DN) center-loss embedding.
- (f) Graph: Calculate 90-ROI functional connectivity and learn the network with (CN) or without (DN) center-loss embedding.
- (f) Graph PCA: Calculate 90-ROI functional connectivity, perform principle component analysis and then learn the network with (CN) or without (DN) center-loss embedding.

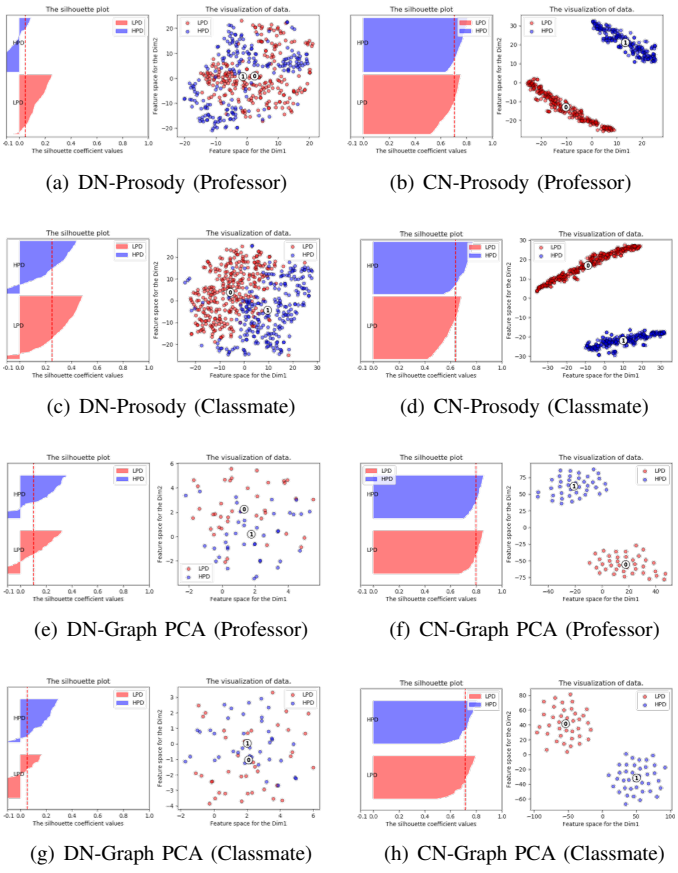
These features are then fed into statistical functional based session-level encoding and further apply decision-level fusion to perform final power distance binary classification for each participant.

B. Experimental Results and Analyses

1) Individual Power Distance Recognition Results: Table II summarizes our experimental results using different modalities in left two columns and associated multimodal fusion results in the right side. There are two different social settings for each subject (P: professor, C: classmate/TA); $P+C$ condition fusion is denoted as the concatenated features from two social settings after CN or DN neural embedding. All evaluation metric used is unweighted average recall (UAR).

Our center-loss embedding network (CN) achieves the best recognition rates among all the modalities, especially in acoustic prosodic network; it obtains 88.5% recognition rate compared to network without center-loss embedding (DN) 61.5%, i.e., 2.7% relative improvement. Further we perform multimodal fusion, which we merge the decision scores of each modality into one feature vector and feed it into final power distance recognition classifier. For instance, $(A)CN + (f)CN - PCA$ indicates concatenation of classifiers' predictions from trained 'audio CN network'

and 'fMRI CN PCA network'. After performing late fusion of audio and fMRI modalities, it reaches the accuracy of 96.2% by integrating CN-Prosody and CN-Graph PCA (7.7% and 23.1% relative improvement over single modality strategy of CN-Prosody and CN-Graph PCA respectively). However, we observe that the fusion of CN-Prosody and CN-PCA negatively impacts the recognition rate even when the CN-PCA achieves a better accuracy than CN-Graph PCA. This might due to the mismatch of computational complexity of prosody and PCA embedding approaches; performing PCA after functional connectivity (Graph) can effectively provide complementary modeling power to acoustic prosodic network. In general, the promising recognition accuracy is attributed to non-linear centralization of both prosodic and neural connectivity features within each social setting, it effectively uncovers the discriminative power of these modalities representation.



(a) DN-Prosody (Professor) (b) CN-Prosody (Professor)

(c) DN-Prosody (Classmate) (d) CN-Prosody (Classmate)

(e) DN-Graph PCA (Professor) (f) CN-Graph PCA (Professor)

(g) DN-Graph PCA (Classmate) (h) CN-Graph PCA (Classmate)

Fig. 3: *The visualization analyses of feature embedding learned with deep network (DN) or center-loss network (CN). Red dots indicates data samples from one of low power distance subject, and Blue dots indicates data samples from one of high power distance subject.*

2) Feature Embedding Visualization: In order to measure the clustering (centralization) effect on the learned feature representations, we first visualize the learned DN-Prosody, CN-Prosody, DN-Graph PCA and CN-Graph PCA features then compute the mean Silhouette Coefficient of all samples

TABLE III: *It summarizes the Silhouette coefficients between low/high power distance and DN/CN models on two social roles and whole group (Avg.).*

		Audio (A)		***fMRI (F)***	
		DN	CN	DN	CN
Professor	LPD	0.1477	0.7114	0.0311	0.9501
	HPD	0.1386	0.6766	0.0703	0.9516
Classmate	LPD	0.1147	0.7655	0.0282	0.9504
	HPD	0.1311	0.7425	0.1089	0.9504
Avg.	LPD	0.1312	0.7385	0.0297	0.9503
	HPD	0.1349	0.7096	0.0896	0.9510

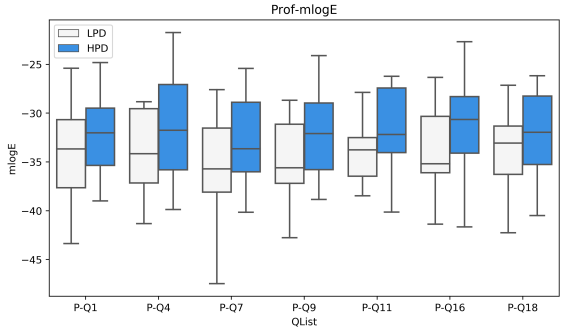
using following equation as measures of clustering effect:

$$s(i) = \frac{a(i) - b(i)}{max\{a(i), b(i)\}} \tag{5}$$

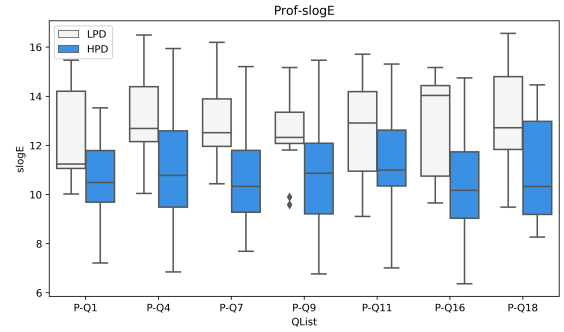
where $a(i)$ is the average distance between point i and other samples in its own class, and $b(i)$ is distance between point i and another class centroid. The $s(i)$ for each sample ranges from -1 (spreading, overlapped cluster) to 1 (tight, well-separated cluster). Here, we report the mean value of Silhouette Coefficient of all samples shown in Table III. In general, we observe that the proposed center network (CN) structure shows a distinct cluster for high and low culture trait of power distance. Without training the framework with center loss criterion, the feature space is highly-overlapping even when the network is tuned using cross entropy with respect to the target power distance label. However, after carrying out center-loss constraint with respect to the social setting, the feature spaces of either of high or low power distance subject are becoming concentrated and clearly non-overlapping on both prosodic and neural connectivity feature spaces. This effect is quite significant when compared between DN and CN networks, especially on fMRI connectivity embedding, i.e., 0.9206 and 0.8614 increased in coefficient values on low and high power distance.

An analysis of visualizing feature embedding under two different social settings (Professor and Classmate) by DN and CN are presented in Figure 3, which shows examples of randomly selected two subjects prosodic and fMRI feature representations without center-loss (DN) and with center-loss embedding (CN) by performing t-SNE visualization. The red dots (cluster 0) indicate data from subject of low power distance, and blue dots (cluster 1) indicate the subject of high power distance. It is clear to observe that the feature spaces after center-loss embedding of either prosodic or fMRI are becoming highly-concentrated and indeed separating each class, showing the effective discriminative modeling ability.

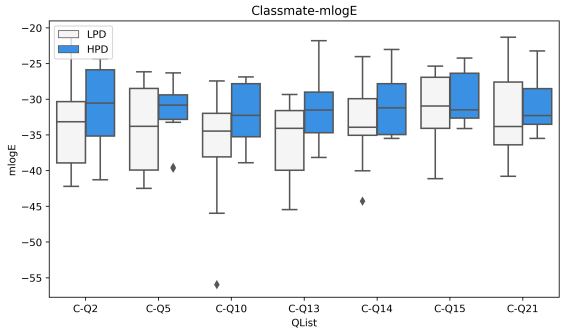
3) Analyses of Acoustic Properties with Power Distance Measure: Since every session is of different length resulting in varying number of sequences, in order to analyze the relationship between the culture power distance and prosodic features intuitively, we compute a total of 38 statistical functional features on $F0$, energy, duration and voice rates, etc; there are then used as the prosodic factors in this analysis. With two social settings, professor and classmate, we calculate the Pearson correlation coefficients between individual power



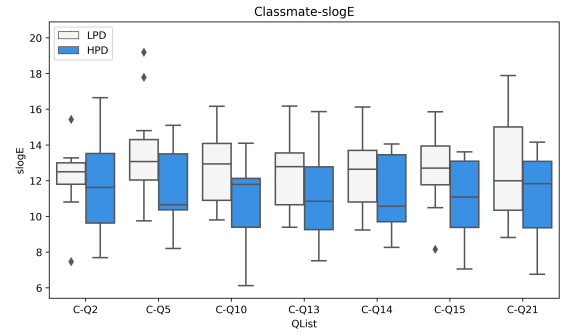
(a) Averaged Energy (Professor)



(b) Std. Energy (Professor)



(c) Averaged Energy (Classmate)



(d) Std. Energy (Classmate)

Fig. 4: *The box-plot analyses between prosodic features and power distance index. The y-axis indicates the values of average value or standard deviation of log energy. The x-axis indicates corresponding question topics of professor or classmate social settings, left and right of each question represent LPD and HPD respectively.*

TABLE IV: *The table summarizes the recognition results that used each ROI region's BOLD signal as feature representation, and lists results that around 0.7 to indicate the discriminative power under regions of interest.*

AAL ROI	UAR (Professor)	UAR (Classmate)	UAR (Fusion)	References
Rolandic operculum (L) ROI-17	0.692	0.654	0.731	Right finger clicking activity during paradigm task
Anterior cingulate (L) ROI-31	0.654	0.692	0.769	[67], [68]
Middle cingulate (R) ROI-34	0.846	0.885	**0.923**	[69]–[71]
Superior occipital (L) ROI-49	0.654	0.692	0.731	Visual attention during paradigm task
Superior occipital (R) ROI-50	0.692	0.731	0.769	Visual attention during paradigm task
Fusiform (L) ROI-56	0.692	0.692	0.731	[72]–[74]
Inferior parietal (L) ROI-61	0.654	0.654	0.692	[75]
Angular (L) ROI-65	0.654	0.654	0.692	[76], [77]
Globus pallidus (R) ROI-76	0.731	0.692	0.769	[78]

distance and each feature to obtain the correlation coefficients. There is an interesting observation from the correlation analysis showing that specifically the voice energy in dB is significantly correlated with an individual power distance measure, i.e., positive correlation with average value of energy in topic 5 of classmate setting, 0.453 ($p = 0.02$), and negative correlation with standard deviation of energy especially in topic 3 of professor setting ($p = 0.003$). The consistent trend of these two prosodic features indicates that indeed prosodic characteristics possess discriminative power for individual status of power distance.

We display box plots of the full range of variation of these two significant prosodics features, i.e., demonstrating minimum to maximum value, the interquartile range (IQR) and the associated typical value (median). There are 7 conversation topics and 2 different social settings displayed in Figure 4, it is clear to notice that averaged energy of HPD are slightly higher than LPD regardless of the social setting. On the contrary, the standard deviation of energy of HPD are slightly lower than LPD no matter in professor or classmate settings, especially obvious in professor setting. It indicates that the distribution of HPD is more concentrated at mean than LPD's for this feature. This may reveal the subjects perceived with high power status tend to have consistent prosodic expressions since lower value of standard deviation.

4) Analyses of Brain Regions with Power Distance Measure: In this section, since we apply Automated Anatomical Labeling (AAL) to divide brain into 116 regions and extract former 90 regions to be our neural connectivity data. In order to investigate regions that impact the performance of individual power distance recognition, we take raw features from BOLD signal to perform power distance recognition using the following steps:

1) extract AAL 90 regions' activation data after fMRI data preprocessing.
2) perform maximum activation value encoding (Max-Pooling) for each session of settings on professor or classmate.
3) perform power distance recognition tasks for professor-setting only, classmate-setting only and fusion of two social settings.

Accuracy is measured using unweighted average recall (UAR) with the evaluation scheme for this task and done via leave-one-person-out cross-validation.

The results are shown on Table IV, we list the UAR around 0.7 of each social setting and also fusion of all social settings. In addition, we also list the ROIs of brains which are correlated to the power distance measure. The most promising accuracy 0.923 obtained is from modeling the right middle cingulate (ROI-34), which outperforms all other ROIs' recognition accuracy. The middle cingulate cortex (MCC) has been identified to support social information processing and boost the execution of social behaviors and processing time of social stimuli [69], [70]. Furthermore, this region is known to play a vital role in performing decision-making tasks which is similar as our block-design fMRI experimental setups. Thus, while these evidences are preliminary, they highlight MCC may be responsible for decision making choices during the situation of social interaction reflecting the differences in PDI value of the subject.

V. Conclusions and Future Works

Culture affects our daily life, influences the manner we learn, live and behave, and further shapes our personalities. Hofstede has developed a theory of cultural dimension that uses six attributes as measures to understand the phenomenon. One of these attributes, power distance index (PDI) is an inner trait that measures an individual's belief about status of power distance which reflects in human's expressive behaviors. While theoretical conceptualization of individual power status belief has received much attention, especially in understanding personalities, limited research has progressed in terms of modeling approaches to automatically recognize one's PDI from either expressive behaviors and/or internal cognitive brain activation. In this work, we present a computational framework in automatically assessing an individual power distance index using a unique dataset with expressive prosodic cues and internal brain connectivity collected from the same subject. Specifically, we propose to learn an enhanced multimodal data representation by jointly considering social settings with center-loss criterion during network training. It successfully achieves an improved recognition accuracy using either single modality or multimodal data from our previous work. By visualizing feature embedding, it reveals that by centralizing the feature representations, our SC-eN effectively enhances the discriminative power of prosodic and neural connectivity representation for power distance recognition. Furthermore, the analyses of acoustic descriptors show an interesting pattern that the voice energy of individuals with HPD is significantly different from LPD. On the other hand, the analyses of fMRI data show right middle cingulate (MCC) plays an important role in information processing of social cognition corroborating with several past research. It points to an evidence that this region may also be responsible for decision-making mechanism for people during interaction revealing different personal power distance attribute.

There are several directions of future works to pursue. First key consideration is to improve the database diversity and scale by seeking partnerships and data from different cultural groups or life experiences and ranks, i.e., teacher vs. student, graduate vs. undergraduate student, and native vs. foreigner. Hofstede indicates the differences of individualism/collectivism, high/low power distance, strong/weak uncertainty avoidance, and masculinity/femininity in teacher/student and student/student interaction [79]. On the other hand, in previous cross-cultural research, [9] recruited a total of 734 participants from China and US, and [80] included 263 workers from 28 different countries to generalize their conclusions. While they have an expanded scale of subjects, none of these works have collected actual behavior data, i.e., mostly rely on self-report measures. Therefore, due to the comparably limited data size in this work, to validate our study further, we would like to perform larger-scale study with additional subjects from different nationalities or with different life experiences or ranks and potentially introduce a more immersive experience for subjects when collecting fMRI data. For example, we could replace the experimental materials used in this work from simply showing text to a first-person view of video stimuli showing an interacting scenario with authoritative professor or TA/classmate. Secondly, we will continue to advance our technical framework to understand the complex interplay between expressive prosodic characteristics when interacting with partners of different power status and internal brain functional activation when performing decision-making during social encounters given this unique collection of datasets.

Furthermore, since almost all studies working on power distance examined it as a cultural factor rather than recognize them as a target through behavior and signal modeling. For example, studies in [8], [9] performed multiple regression to examine whether power distance influence cultures of different nationalities. By having a data-driven learning-based analytics in quantifying cultural construct of an individual power distance, we would continue to expand our research scale with different disciplines across nations to have a more objective method in deeper understanding of cross-cultural phenomenon. Finally, we would like to continue understand whether other dimensions of culture attributes could also be inferred automatically through the combination of expressive behavior cues and internal brain functions to further enable better design of human-centered technology and services for real world applications.

REFERENCES

[1] E. B. Tylor, *Primitive culture: researches into the development of mythology, philosophy, religion, art, and custom*. J. Murray, 1871, vol. 2.

[2] F. M. Keesing, *Cultural anthropology*. Mittal Publications, 1963.

[3] A. Bandura, "Social cognitive theory in cultural context," *Applied psychology*, vol. 51, no. 2, pp. 269–290, 2002.

[4] L. K. Trevino, "Ethical decision making in organizations: A person-situation interactionist model," *Academy of management Review*, vol. 11, no. 3, pp. 601–617, 1986.

[5] G. Hofstede, G. J. Hofstede, and M. Minkov, *Cultures and organizations: Software of the mind*. Citeseer, 2005, vol. 2.

[6] G. Hofstede, *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001.

[7] V. Taras, B. L. Kirkman, and P. Steel, "Examining the impact of culture's consequences: A three-decade, multilevel, meta-analytic review of hofstede's cultural value dimensions." *Journal of applied psychology*, vol. 95, no. 3, p. 405, 2010.

[8] J. Brockner, G. Ackerman, J. Greenberg, M. J. Gelfand, A. M. Francesco, Z. X. Chen, K. Leung, G. Bierbrauer, C. Gomez, B. L. Kirkman *et al.*, "Culture and procedural justice: The influence of power distance on reactions to voice," *Journal of Experimental Social Psychology*, vol. 37, no. 4, pp. 300–315, 2001.

[9] B. L. Kirkman, G. Chen, J.-L. Farh, Z. X. Chen, and K. B. Lowe, "Individual power distance orientation and follower reactions to transformational leaders: A cross-level, cross-cultural examination," *Academy of Management Journal*, vol. 52, no. 4, pp. 744–764, 2009.

[10] G. Hofstede, "Cultural dimensions in management and planning," *Asia Pacific journal of management*, vol. 1, no. 2, pp. 81–99, 1984.

[11] A. Jaimes and N. Dimitrova, "Human-centered multimedia: culture, deployment, and access," *IEEE MultiMedia*, vol. 13, no. 1, pp. 12–19, 2006.

[12] L. Batrinca, N. Mana, B. Lepri, N. Sebe, and F. Pianesi, "Multimodal personality recognition in collaborative goal-oriented tasks," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 659–673, 2016.

[13] J.-I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 41–55, 2012.

[14] M. J. Scott, S. C. Guntuku, W. Lin, and G. Ghinea, "Do personality and culture influence perceived video quality and enjoyment?" *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1796–1807, 2016.

[15] M. J. Scott, S. C. Guntuku, Y. Huan, W. Lin, and G. Ghinea, "Modelling human factors in perceptual multimedia quality: On the role of personality and culture," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 481–490.

[16] S. C. Guntuku, W. Lin, M. J. Scott, and G. Ghinea, "Modelling the influence of personality and culture on affect and enjoyment in multimedia," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 236–242.

[17] P. Varini, G. Serra, and R. Cucchiara, "Personalized egocentric video summarization of cultural tour on user preferences input," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2832–2845, 2017.

[18] R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, and V. Gupta, *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications, 2004.

[19] S. H. Schwartz, "Beyond individualism/collectivism: New cultural dimensions of values." 1994.

[20] P. B. Smith, S. Dugan, and F. Trompenaars, "National culture and the values of organizational employees: A dimensional analysis across 43 nations," *Journal of cross-cultural psychology*, vol. 27, no. 2, pp. 231–264, 1996.

[21] V. Taras, P. Steel, and B. L. Kirkman, "Improving national cultural indices using a longitudinal meta-analysis of hofstede's dimensions," *Journal of World Business*, vol. 47, no. 3, pp. 329–341, 2012.

[22] C. Demetriou, B. U. Ozer, and C. A. Essau, "Self-report questionnaires," *The encyclopedia of clinical psychology*, pp. 1–6, 2014.

[23] D. Stevanovic, R. Urbán, O. Atilola, P. Vostanis, Y. S. Balhara, M. Avicenna, H. Kandemir, R. Knez, T. Franic, and P. Petrov, "Does the strengths and difficulties questionnaire–self report yield invariant measurements across different nations? data from the international child mental health study group," *Epidemiology and psychiatric sciences*, vol. 24, no. 4, p. 323, 2015.

[24] V. A. Scholtes, C. B. Terwee, and R. W. Poolman, "What makes a measurement instrument valid and reliable?" *Injury*, vol. 42, no. 3, pp. 236–240, 2011.

[25] R. W. Picard, *Affective computing*. MIT press, 2000.

[26] C. M. Lee, S. S. Narayanan *et al.*, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[27] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[28] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[29] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 341–359, 2013.

[30] E. Crane and M. Gross, "Motion capture and emotion: Affect detection in whole body movement," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 95–101.

[31] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[32] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1175–1191, 2001.

[33] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[34] H. Gunes, M. Piccardi, and M. Pantic, "From the lab to the real world: Affect recognition using multiple cues and modalities," in *Affective Computing*. IntechOpen, 2008.

[35] F.-S. Tsai, H.-C. Yang, W.-W. Chang, and C.-C. Lee, "Automatic assessment of individual culture attribute of power distance using a social context-enhanced prosodic network representation." in *Interspeech*, 2018, pp. 436–440.

[36] V. Aubergé, "A gestalt morphology of prosody directed by functions: the example of a step by step model developed at icp," in *Speech Prosody 2002, International Conference*, 2002.

[37] A. K. Uskul, S. Paulmann, and M. Weick, "Social power and recognition of emotional prosody: High power is associated with lower recognition accuracy than low power." *Emotion*, vol. 16, no. 1, p. 11, 2016.

[38] M. Weick, A. Guinote, and D. Wilkinson, "Lack of power enhances visual perceptual discrimination." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 65, no. 3, p. 208, 2011.

[39] H. Mixdorff, A. Hönemann, and A. Rilliard, "Acoustic-prosodic analysis of attitudinal expressions in german," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[40] R. Van Bezooijen and C. Gooskens, "Identification of language varieties: The contribution of different linguistic levels," *Journal of language and social psychology*, vol. 18, no. 1, pp. 31–48, 1999.

[41] D. S. Hurley, "Issues in teaching pragmatics, prosody, and non-verbal communication," *Applied Linguistics*, vol. 13, no. 3, pp. 259–280, 1992.

[42] J. West and J. L. Graham, "A linguistic-based measure of cultural distance and its relationship to managerial values," *MIR: Management International Review*, pp. 239–260, 2004.

[43] S. Grawunder, M. Oertel, and C. Schwarze, "Politeness, culture, and speaking task–paralinguistic prosodic behavior of speakers from austria and germany," in *Speech Prosody*, vol. 2014, 2014, pp. 159–163.

[44] A. Barbulescu, R. Ronfard, and G. Bailly, "Which prosodic features contribute to the recognition of dramatic attitudes?" *Speech Communication*, vol. 95, pp. 78–86, 2017.

[45] T. Shochi, A. Rilliard, and V. Aubergé, "Donna erickson intercultural perception of english, french and japanese social affective prosody," *The role of prosody in Affective Speech*, vol. 97, p. 31, 2009.

[46] J. E. Koski, H. Xie, and I. R. Olson, "Understanding social hierarchies: The neural and psychological foundations of status perception," *Social neuroscience*, vol. 10, no. 5, pp. 527–550, 2015.

[47] S.-L. Liew, Y. Ma, S. Han, and L. Aziz-Zadeh, "Who's afraid of the boss: cultural differences in social hierarchies modulate self-face recognition in chinese and americans," *PloS one*, vol. 6, no. 2, p. e16901, 2011.

[48] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A survey of autonomous human affect detection methods for social robots engaged in natural hri," *Journal of Intelligent & Robotic Systems*, vol. 82, no. 1, pp. 101–133, 2016.

[49] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language." *Journal of personality and social psychology*, vol. 108, no. 6, p. 934, 2015.

[50] M. B. Harms, A. Martin, and G. L. Wallace, "Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies," *Neuropsychology review*, vol. 20, no. 3, pp. 290–322, 2010.

[51] J. A. Mumford, "A power calculation guide for fmri studies," *Social cognitive and affective neuroscience*, vol. 7, no. 6, pp. 738–742, 2012.

[52] D. Szucs and J. P. Ioannidis, "Sample size evolution in neuroimaging research: an evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals," *NeuroImage*, p. 117164, 2020.

[53] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[54] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[55] C.-Y. Lin and H.-C. Wang, "Language identification using pitch contour information," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–601.

[56] J. M. Shine, O. Koyejo, and R. A. Poldrack, "Temporal metastates are associated with differential patterns of time-resolved connectivity, network topology, and attention," *Proceedings of the National Academy of Sciences*, vol. 113, no. 35, pp. 9888–9891, 2016.

[57] D. S. Bassett, M. Yang, N. F. Wymbs, and S. T. Grafton, "Learning-induced autonomy of sensorimotor systems," *Nature neuroscience*, vol. 18, no. 5, p. 744, 2015.

[58] H. You, A. Liska, N. Russell, and P. Das, "Automated brain state identification using graph embedding," in *2017 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. IEEE, 2017, pp. 1–5.

[59] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[60] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. OReilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 302–309.

[61] M. Jiang, Z. Yang, W. Liu, and X. Liu, "Additive margin softmax with center loss for face recognition," in *Proceedings of the 2018 the 2nd International Conference on Video and Image Processing*. ACM, 2018, pp. 1–6.

[62] Y. Xu, H. Ma, L. Cao, H. Cao, Y. Zhai, V. Piuri, and F. Scotti, "Robust face recognition based on convolutional neural network," *DEStech Transactions on Computer Science and Engineering*, no. icmsie, 2017.

[63] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. OReilly, S. Han, P. Liu, M. Chen, and Y. Tong, "Feature-level and model-level audiovisual fusion for emotion recognition in the wild," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019, pp. 443–448.

[64] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7405–7409.

[65] R. Zhang, Q. Wang, and Y. Lu, "Combination of resnet and center loss based metric learning for handwritten chinese character recognition," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 5. IEEE, 2017, pp. 25–29.

[66] S. Yang, F. Nian, and T. Li, "A light and discriminative deep networks for off-line handwritten chinese character recognition," in *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, 2017, pp. 785–790.

[67] N. I. Eisenberger and M. D. Lieberman, "Why rejection hurts: a common neural alarm system for physical and social pain," *Trends in cognitive sciences*, vol. 8, no. 7, pp. 294–300, 2004.

[68] M. A. Apps, M. F. Rushworth, and S. W. Chang, "The anterior cingulate gyrus and social cognition: tracking the motivation of others," *Neuron*, vol. 90, no. 4, pp. 692–707, 2016.

[69] K. Hadland, M. F. Rushworth, D. Gaffan, and R. Passingham, "The effect of cingulate lesions on social behaviour and emotion," *Neuropsychologia*, vol. 41, no. 8, pp. 919–931, 2003.

[70] P. H. Rudebeck, M. J. Buckley, M. E. Walton, and M. F. Rushworth, "A role for the macaque anterior cingulate gyrus in social valuation," *Science*, vol. 313, no. 5791, pp. 1310–1312, 2006.

[71] M. A. Apps, P. L. Lockwood, and J. H. Balsters, "The role of the midcingulate cortex in monitoring others' decisions," *Frontiers in neuroscience*, vol. 7, p. 251, 2013.

[72] R. Adolphs, "The neurobiology of social cognition," *Current opinion in neurobiology*, vol. 11, no. 2, pp. 231–239, 2001.

[73] B. Rossion, R. Caldara, M. Seghier, A.-M. Schuller, F. Lazeyras, and E. Mayer, "A network of occipito-temporal face-sensitive areas besides the right middle fusiform gyrus is necessary for normal face processing," *Brain*, vol. 126, no. 11, pp. 2381–2395, 2003.

[74] H. D. Critchley, E. M. Daly, E. T. Bullmore, S. C. Williams, T. Van Amelsvoort, D. M. Robertson, A. Rowe, M. Phillips, G. McAlonan, P. Howlin *et al.*, "The functional neuroanatomy of social behaviour: changes in cerebral blood flow when people with autistic disorder process facial expressions," *Brain*, vol. 123, no. 11, pp. 2203–2212, 2000.

[75] J. Y. Chiao, T. Harada, E. R. Oby, Z. Li, T. Parrish, and D. J. Bridge, "Neural representations of social status hierarchy in human inferior parietal cortex," *Neuropsychologia*, vol. 47, no. 2, pp. 354–363, 2009.

[76] A. L. Glenn, A. Raine, and R. A. Schug, "The neural correlates of moral decision-making in psychopathy," *Molecular psychiatry*, vol. 14, no. 1, p. 5, 2009.

[77] D. Bzdok, G. Hartwigsen, A. Reid, A. R. Laird, P. T. Fox, and S. B. Eickhoff, "Left inferior parietal lobe engagement in social cognition and language," *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 319–334, 2016.

[78] L. A. McGraw and L. J. Young, "The prairie vole: an emerging model organism for understanding the social brain," *Trends in neurosciences*, vol. 33, no. 2, pp. 103–109, 2010.

[79] G. Hofstede, "Cultural differences in teaching and learning," *International Journal of intercultural relations*, vol. 10, no. 3, pp. 301–320, 1986.

[80] S. Bochner and B. Hesketh, "Power distance, individualism/collectivism, and job-related attitudes in a culturally diverse work group," *Journal of cross-cultural psychology*, vol. 25, no. 2, pp. 233–257, 1994.

Fu-Sheng Tsai (Student Member, IEEE) is working toward the Ph.D. degree in the Electrical Engineering Department, NTHU, Hsinchu Taiwan. He received his B.S. degree in Communication Engineering from the National Taiwan Ocean University (NTOU), Taiwan in 2015. His research interests are in human-centered behavioral signal processing (BSP), focusing on development of multimodal signal processing for pain and affect computation. He is also a Student Member of the IEEE Signal Processing Society.

Wei-Wen Chang is a Professor of the International Human Resource Development at the National Taiwan Normal University, Taiwan. She received her Ph.D. in Adult Learning from University of Wisconsin-Madison. She studies multinational HRD, intercultural training, and cross-cultural competency, and her work has been published in journals such as the International Journal of Intercultural Relations, Nonprofit and Voluntary Sector Quarterly, Adult Education Quarterly, Human Resource Development Quarterly, Human Resource Development International, and Human Resource Development Review.

Chi-Chun Lee (M'13, S'20) is an Associate Professor at the Department of Electrical Engineering with joint appointment at the Institute of Communication Engineering of the National Tsing Hua University (NTHU), Taiwan. He received his B.S. and Ph.D. degree both in Electrical Engineering from the University of Southern California, USA in 2007 and 2012. His research interests are in speech and language, affective multimedia, health analytics, and behavior computing. He is an associate editor for the IEEE Transaction on Affective Computing (2020-), the IEEE Transaction on Multimedia (2019-2020), and a TPC member for APSIPA IVM and MLDA committee. He serves as an area chair for INTERSPEECH 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity chair for ACM ICMI 2018, sponsorship and special session chair for ISCSLP 2018, 2020, and a guest editor in Journal of Computer Speech and Language on special issue of Speech and Language Processing for Behavioral and Mental Health.

He is the recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), the CIEE Outstanding Young Electrical Engineer Award (2020), the IICM K. T. Li Young Researcher Award (2020), the MOST Futuretek Breakthrough Award (2018, 2019). He led a team to the 1st place in Emotion Challenge in INTERSPEECH 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in INTERSPEECH 2019. He is a coauthor on the best paper award/finalist in INTERSPEECH 2008, INTERSPEECH 2010, IEEE EMBC 2018, INTERSPEECH 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in Journal of Speech Communication. He is an IEEE senior member and a ACM and ISCA member