Received November 2, 2020, accepted December 4, 2020, date of publication December 14, 2020,
date of current version December 30, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3044285*

# Prediction of Inter-Personal Trust and Team Familiarity From Speech: A Double Transfer Learning Approach

**CATHERINE SANDOVAL RODRIGUEZ[1], (Graduate Student Member, IEEE), APRIL ROSE PANGANIBAN[2], MELISSA N. STOLAR[3], ROBERT BOLIA[3], AND MARGARET LECH[1], (Member, IEEE)**

[1]School of Engineering, RMIT University, Melbourne, VIC 3000, Australia
[2]B CIV USAF AFMC 711 HPW/RHXS, Wright-Patterson Air Force Base, Dayton, OH 45433, USA
[3]Defence Science and Technology Group, Aerospace Division, Fishermans Bend, VIC 3207, Australia

Corresponding author: Margaret Lech (margaret.lech@rmit.edu.au)

This work was supported in part by the Defence Science Institute Postgraduate Scholarship, and in part by the Commonwealth of Australia as represented by the Defence Science and Technology (DST) Group under Research Agreement MyIP7098.

**ABSTRACT** Speech classification is one of the most convenient objective measures of internal state exhibited during a problem-solving task that requires verbal communication. This study investigates the hypothesis of speech acoustic characteristics being indicative of trust between team members and team members' familiarity with each other. Speech recordings from 27 dyadic teams (26 males and 28 females) were made during a distributed threat perception task, determining safe points along a route through the town to be visited by a VIP. Before the threat detection mission, 26 team members knew each other, and the remaining 28 had no prior knowledge of their partners. Two levels (Low Trust and High Trust) of two trust constructs, TTP (Trust, Trustworthiness, Propensity to trust), and RIS (Reliance Intentions Scale), were estimated based on numerical responses to pre- and post-mission surveys. Speech recordings of individual speakers were divided into 1-second intervals and converted into RGB images of amplitude spectrograms. The images were classified using a pre-trained convolutional neural network ResNet-18 fine-tuned to recognize either the trust level or familiarity. In the baseline classification scenario, the speech was classified using a single transfer learning into Low/High-trust categories separately for RIS and TTP constructs before and after the mission yielding an average classification accuracy of 82%-86%. Single transfer learning classification into Know/Unknown-partners categories led to 85% accuracy. Application of double transfer learning, i.e., first tuning the ResNet-18 on Know/Unknown labels and then on Low/High-trust, increased the trust classification accuracy up to 89%. When tuning the ResNet-18 on Low/High-trust and then on Known/Unknown labels, the accuracy of partner familiarity recognition was also increased up to 89%. These results support the hypothesis of speech acoustics being indicative of trust and familiarity between team members and show that by adding prior related knowledge to the model, more efficient learning can be achieved without increasing the training data size.

**INDEX TERMS** Inter-personal trust prediction, inter-personal familiarity prediction, speech classification, transfer learning, convolutional neural networks.

## I. INTRODUCTION

Interpersonal trust is commonly defined as the willingness to accept vulnerability to another person's actions or decisions. Trust is a vital component of daily human interactions and decision-making processes with a partner. Interpersonal trust has high importance in all aspects of life. Understanding of subjective and objective factors affecting trust is an active topic of research in psychology, social studies, and recently in artificial intelligence and machine learning. A comprehensive review of the interpersonal trust research from the perspective of behavioral psychology can be found in [1]. Recently

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

investigated objective trust indicators include emotional states [2], facial expressions [3], and speech attributes [4]. From the machine learning perspective, speech as an objective factor affecting interpersonal trust is of particular interest. Through the mapping of social attributes of trustworthiness into synthetic speech and machine-made decision-making processes, higher social acceptance of all kinds of human-machine interactions can be achieved. Synthetic speech produced by robots can be more friendly when characterized by acoustic and verbal signs of empathy, affection, and trustworthiness. Verbal human-machine communications containing acoustic traits of trust are vital for achieving high performing human-machine teams working on high-responsibility tasks of public security and defense.

This paper investigates the hypothesis of acoustic characteristics of speech being indicative of inter-personal trust and team members' familiarity with each other. Two fundamental research questions underlie this study. Firstly, we wanted to find out if it is possible to predict the trust and familiarity of team members working on a distributed problem-solving task from the acoustic characteristics of their speech. Secondly, we wanted to determine if having a pre-requisite knowledge of familiarity can improve the prediction of trust and vice versa.

In general, there are two types of trusters [3]: individuals who do not trust others until they gather clear evidence that the trust can be granted and individuals who exhibit a high degree of prior trust in others until they obtain evidence that their trust cannot be granted anymore. These two groups were identified in this work by conducting pre- and post-mission surveys determining how much team members trust each other. Given the two trust labels (pre- and post-mission), an automatic prediction of pre- and post-missional trust was conducted using speech collected from team members during the problem-solving mission.

The remaining parts of the paper are organized as follows. Section II provides an overview of related studies. Section III describes the speech data and the methodology of speech classification. Section IV describes the experiments and discusses the results. Section V concludes the paper.

## II. RELATED WORKS

Trust has been analyzed from the perspective of several disciplines, including sociology, psychology, philosophy, management, economics, automation, and communication networks. In general, trust can be defined as a relationship between two entities denominated trustor (evaluator) and trustee (evaluate), based on a given criterion [5]. Trust is a multidisciplinary concept that can be applied to human-human interactions as well as to other types of relationships, such as human-machine interactions.

Although there is neither a multidisciplinary trust model nor a standard metric for the quantification of trust measurements, there are some common tendencies in trust investigations. In particular, there is a recently growing interest in studies of trust between humans and artificial agents (human-robot interactions). As a result, several trust models have been proposed, which can have an online or offline structure [6]. In behavioral economics, the "*Trust Game*," designed in [7], is commonly applied to evaluate the level of trust in economic decisions. One of the most widely used trust models was developed by Mayer *et al.* [8] in the context of organizational management. This model defines trust using three dimensions: ability, benevolence, and integrity (ABI framework). The gender rules of interpersonal trust based on a research questionnaire [9] revealed that male college students have more difficulty trusting other students than female college students. In a search for objective predictors of trust, facial expressions have been considered as a guide to trustworthiness. It has been observed that physical facial attributes, as well as the degree of similarity to oneself, can affect the perceived trustworthiness of a person. However, the perception of acoustical qualities which might establish trust are easier to obtain for remote teams and individuals. In general, people more similar to oneself are perceived as more trustworthy [10]. In the absence of cues indicating similarity, as is the case in our current world of remote work and teleconferencing, increased pitch and intensity are perceived as indicators of deception or untrustworthiness [11]. However, as described further, trust perception in speech has many viable indicators and what may be asked is the content of the shared speech. There is a difference between intended deception and determining the trustworthiness of a partner who harbors no malintent.

Several studies have investigated the relationship between trust perception and speech [12]–[16]. In [12], voice models were evaluated to investigate the correlation between basic acoustical parameters and the perception of low and high trust. Results showed that a significant change in intonation is related to an increase in the perception of trust. Similar studies conducted in [13] found an inverse correlation between vocal pitch and trust.

Experiments conducted in [14] reported that low harmonic-to-noise ratio, low mean fundamental frequency, and a fast speech rate are the most significant acoustic characteristics present in the perception of high trust. Additionally, tests comparing positive valence speakers against negative valence ones, younger voices against those with older voices, and female against male speakers showed that a higher perception of trust is achieved for positive valence speakers, younger voices, and female speakers, respectively.

Recently, studies conducted in [15] investigated how the perception of trust is affected by the voice and accent of a speaker. Trust was evaluated as if a statement is believable or not, using confident and doubtful voices. The study was conducted from the perspective of an English Canadian listener. Three accent groups of speakers were established: "*in a group*" (native Canadian English speakers), "*out group-regional*" (Australian English accent), and "*out-group/foreign*" (Canadian French accent). The results indicated that the accent does not affect favorable trust in confident statements against doubtful ones. However, confident

statements produced by the out-group/foreign accent were evaluated as less believable than the statements produced by the English accent groups. Evaluation of basic acoustic features showed that doubtful statements present lower speaker rates and amplitude ranges than confident expressions in all accents.

Acoustic speech properties of high and low public trust politicians have been recently investigated in [4]. Statistical analysis and classification of low-level acoustic speech parameters revealed strong gender differences. In general, high trust females appeared to have higher spectral energy for both voiced and unvoiced speech within 0 - 500 Hz compared to low trust females (sonorant sound). It was also revealed that trust in females is related to the timber and non-verbal attributes of speech that are often linked to emotions [17], [18]. High trust males, on the other hand, had slightly lower energy concentration at low frequencies (more fricative or constrained sound) compared to low trust males. The main differences between spectral energies of high and low trust males were found to be in the unvoiced speech within the low-frequency band 500 – 1500 Hz. High trust males appear to have a smaller spectral slope within this range (i.e., smoother transitions between slowly and fast-changing speech components). In addition, high trust males showed larger spectral flux (i.e., larger differences between spectra of consecutive frames). An automatic speech classification into low and high-trust politicians using a multilayer perceptron neural network led to an average accuracy of 81%. The outcomes of this study indicated that the acoustic speech properties of politicians are good indicators of public trust.

The recent advent of deep learning (DL) neural network techniques have been particularly helpful in creating very efficient prediction models. The availability of cloud storage spaces and graphical processing units (GPUs) provide support for dealing with large numbers of data and vast numbers of computations needed to train the DL models. Classical signal processing pipelines, including complex pre-processing of speech followed by feature extraction and feature selection, were simplified by application end-to-end neural network approaches. Convolutional neural network (CNN) models, for example, calculate their own features through the iterative generation of correlated data and thus dropping the need to compute problem-related features. Available pre-trained neural network models reduce the computational and data requirements allowing to perform specific problem-related fine-tuning on a relatively small number of available data. The network's features, also known as network embeddings, provide a highly competitive alternative to handcrafted features for classical classifiers such as the support vector machine (SVM), k-nearest neighbors (KNN), or the Gaussian mixture model (GMM).

Although automated analyses of social signals can provide supplementary information and discover trends that human beings are not able to detect, machine learning techniques for social signal processing have been rarely used in the studies of human-human trust interaction. An approach to an automatic prediction of trust in human-human interaction was presented in [19]. The investigation did not involve speech analysis; only features of the non-verbal behavior of the participant (movement of the head, arms, eyes, touch, and smile signals) were analyzed. Hidden Markov models were implemented to investigate the temporal relationships between trust and social signals. Trust in human-machine-human interactions was analyzed in [20], [21] using an automated approach that combines subjective labels and physical features. Two scenarios were evaluated using two scripted dialogues. The first dialogue was video recorded during a business conversation, and the second one during a fire rescue. A multimodal analysis was performed using facial expressions, basic acoustic features (fundamental frequency, voice/silence ratio, mean voice power, and energy entropy), and the heart rate variability. An ensemble classification was then applied to predict three levels of trust using a neural network (NN) with a fuzzy decision as a classifier.

In this study, we provide a twofold contribution to automatic trust recognition from speech using convolutional neural networks (CNNs). Firstly, we demonstrate that both interpersonal trust between team members and team members' familiarity with each other can be efficiently predicted from speech. Secondly, we show that a transfer of team familiarity and inter-personal trust knowledge between deep CNN models can enhance the models' performance.

## III. METHOD

### A. SPEECH DATA

#### 1) DATA COLLECTION

The speech data was collected at the Air Force Research Laboratory (AFRL) from dyadic team partners working on a distributed problem-solving task aided by a computer video simulation program. The task was to approve the safety of the travel route for a VIP. During the task, team members were conducting verbal communication, and their speech was recorded. When working on the task, team members were required to make a few intermediate decisions by weighing their own information against information obtained from the team partner. Since the information available to each team member was different, a certain amount of trust or mistrust in the partner was required from team members to make the decisions.

#### 2) DATA LABELING

Participants were asked to provide responses to pre- and post-mission psychology surveys with the same questions used in both. The survey responses were used to evaluate two trust-related construct measures called TTP (Trust, Trustworthiness, Propensity to trust), and RIS (Reliance Intentions Scale) [22]. Table 1 shows the questions used to derive the TTP trust construct measure. The response to each TTP question was given by the participants as an integer number on a scale from 1 to 5. PreTTP and PostTTP scores were calculated

**TABLE 1. TTP Trust Construct Questions. Answers on Scale 0-5.**

| Index | TTP Trust Questions |
|---|---|
| A | If I had my way, I wouldn't let My partner have any influence over issues that are important to me. |
| B | I would be willing to let My partner have complete control over my future of this study. |
| C | I really wish I had a good way to keep an eye on My partner. |
| D | I would be comfortable giving My partner a task or problem which was critical to me, even if I could not monitor their actions. |

**TABLE 2. RIS Construct Questions. Answers on Scale 0-7.**

| Index | RIS Questions |
|---|---|
| A | I would rely on my partner without hesitation. |
| B | I think working with my partner will lead to positive outcomes. |
| C | I would feel comfortable relying on my partner in the future. |
| D | When the task was hard, I felt like I could depend on my partner. |
| E | If I were facing a very hard task in the future, I would want to have my partner with me. |
| F | I would be comfortable allowing my partner to make all decisions. |
| G | If I had it my way, I would NOT let my partner have any influence over issues that are important to the task(surveillance). |
| H | I would be comfortable giving my partner complete responsibility for the surveillance task. |

for each participant, as

$$\text{PreTTP} = \text{PostTTP} = \frac{A + B + (6 - C) + D}{4} \quad (1)$$

where A, B, C, and D denote pre- or post-mission scores given by the participants in response to questions listed in Table 1. Speech of individuals that obtained PreTTP and PostTTP scores above the average estimated across all participants was labeled as ''TTP*HighTrust,*'' and speech representing scores falling below this average was labeled as ''TTP*LowTrust*.''

A similar speech labeling procedure was applied to the RIS trust construct. However, it this case, the response to each RIS question was given by the participants as an integer number on a scale from 1 to 7, and the PreRIS and PostRIS scores were calculated for each participant, as

$$\begin{aligned}\text{PreRIS} &= \text{PostRIS} \\ &= \frac{A + B + C + D + E + F + (8 - G) + H}{8}\end{aligned} \quad (2)$$

where A, B, C, D, E, F, G, and H denote pre- or post-mission scores given by the participants in response to questions listed in Table 2. Speech of individuals that obtained PreRIS and PostRIS scores above the average estimated across all participants was labeled as ''RIS*HighTrust*'' and speech representing scores falling below this average was labeled as ''RIS*LowTrust*.''

#### 3) DATA CHARACTERISTICS

Speech recordings were collected from 27 dyadic teams (i.e., 54 individuals) participating in the problem-solving task. The data distribution across genders and team members' familiarity with each other is summarized in Table 3. The term ''*Known*'' indicates participants who knew their partners before the data collection, and ''*Unknown*'' refers to participants who did not know their partners before the mission. The duration of audio recordings collected from each participant was 10-15 minutes.

Fig. 1 shows the data distribution across two different trust levels (High/Low) and participant familiarity

**TABLE 3. Data Distribution Across Genders And Team Members Familiarity.**

| Total No of Participants: 56 (100%) | | | |
|---|---|---|---|
| Male | | Female | |
| 26 (48%) | | 28 (52%) | |
| Known | Unknown | Known | Unknown |
| 12 | 14 | 14 | 14 |

(Known/Unknown) in four cases (PreTTP, PostTTP, PreRIS, and PostRIS). It shows apparent differences in trust assessment between pre and post-missional surveys, indicating that the problem-solving task had an effect on inter-personal trust between team members. The trust assessment also differs between TTP and RIS surveys and depends on team members' familiarity. Given these observations, we wanted to determine if these differences can be predicted through an automatic speech classification. To do so, we conducted speech classification experiments on the following binary categories (*Low* vs. *High* trust), team familiarity (*Known vs. Unknown*), and trust change during the problem-solving task (*Change* vs. *No Change*). In addition, we have investigated a multilabel classification combining two labels (trust and team familiarity) and three labels (trust, team familiarity, and trust change). Since the team familiarity and interpersonal trust are likely to be correlated, we also wanted to know if a classification system capable of familiarity prediction can be trained more efficiently to determine trust compared to a system trained only to predict trust without prior knowledge of familiarity and vice versa.

### B. SPEECH CLASSIFICATION METHOD

#### 1) SPEECH CLASSIFICATION FRAMEWORK

Given the limited size of available data and computational resources, the option of training a designated trust-prediction CNN ''*from scratch*'' was not feasible. We have, therefore, adopted a classification approach proposed in [23] for the prediction of speech emotions. However, instead of AlexNet, the more advanced convolutional neural network (CNN)

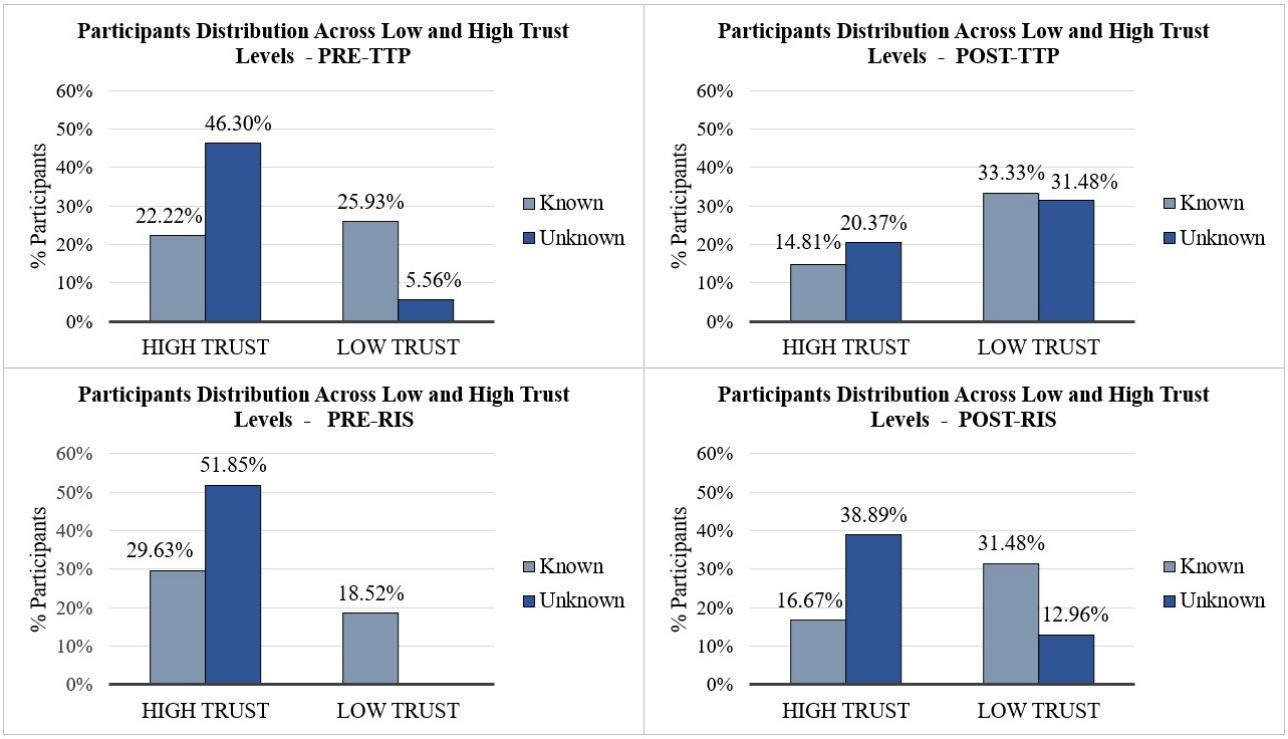


**FIGURE 1.** Participants characteristics based on responses to TTP and RIS questions pre- and post-experiment. The term known indicates team members who knew each other before the experiment, and the term unknown refers to team members that did not know each other before the experiment.

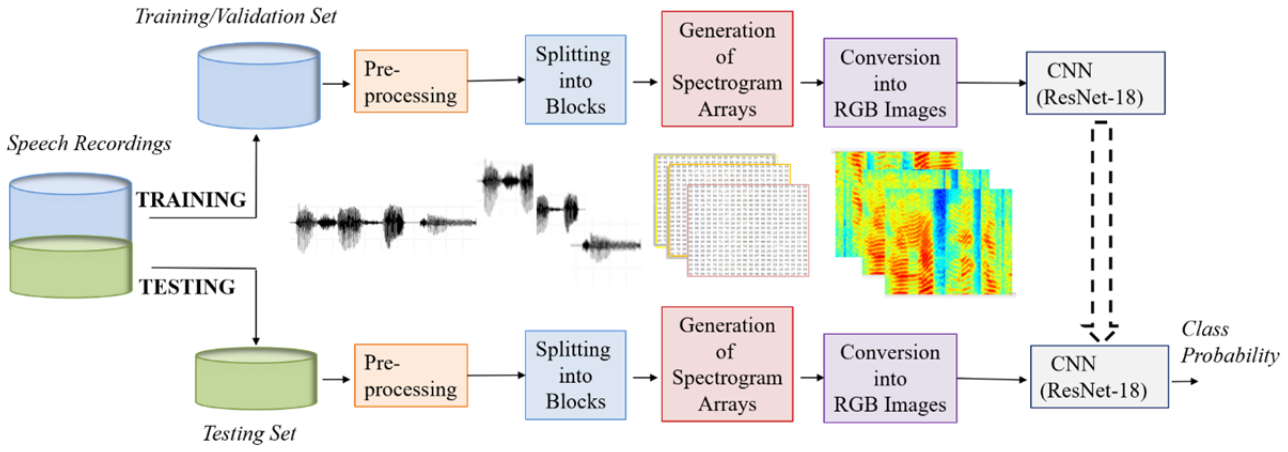


**FIGURE 2.** CNN model training and testing framework.

structure of ResNet-18 [24] was applied. Fig. 2 shows an overview of the classification procedure consisting of the training and the testing stages. The following paragraphs provide brief descriptions of each processing step shown in Fig. 2.

### 2) PRE-PROCESSING

The pre-processing step included speech normalization into the range $-1$ to $+1$ and removal of silence intervals. The silence was detected using an empirically chosen energy threshold. No speech pre-emphasis filter was applied. Since the signal was recorded from personal headset microphones within a quiet room, the signal to noise ratio was relatively high, and there was no need for the removal of noise or any other interference. The sampling rate was 16 kHz giving the 8 kHz signal bandwidth.

### 3) SPLITTING INTO BLOCKS

Pre-processed waveforms of speech samples were divided into short 1-second blocks to conduct block-by-block

spectrogram computation. A short, 10-millisecond stride was applied between subsequent blocks. For each block, a single spectral amplitude array was calculated and converted into a color RGB image. Waveforms remainders that did not fill up to 1-second frame length were discarded. The 1-second block-duration of 1 second was consistent with the previously reported duration used in speech-based prediction of speaker's states [24]–[26]. The stride time between subsequent blocks was chosen in an arbitrary way. By changing the stride duration, a smaller or larger number of CNN input images could be generated. The 10-millisecond stride allowed to generate a relatively large number of training images ensuring sufficient generalization of the CNN model.

#### 4) GENERATION OF SPECTROGRAM IMAGES

A short-time Fourier transform was performed on each 1-second block of speech waveforms. The resulting pairs of real and imaginary outputs were converted into real-valued spectral magnitudes and concatenated across all subsequent frames within a given block to form a time-frequency spectral magnitude array of 257 x 259 real-valued numbers representing a given block. Where 257 was the number of discrete-frequency values, and 259 was the number of discrete-time values for which the spectrogram arrays were evaluated. The time scale of the magnitude spectrograms was linear, spanning the range from 0 to 1 second. The frequency scale, on the other hand, was logarithmic, spanning the entire bandwidth range from 0 to 8 kHz. The logarithmic frequency scale was previously reported to give superior performance compared to linear, mel, and ERB scales [16], [20].

The advantage of using the logarithmic scale is that compared to the other scales, it compresses the high-frequency details and expand the low-frequency range where the vital information regarding the speaker's fundamental frequency and the first formant values is given.

Spectral magnitude arrays were converted into the RGB image format represented by three color-component arrays. As shown in [23], the RGB images of amplitude spectrograms provide higher speech classification performance compared to the greyscale images or raw spectrogram arrays. This is largely due to the fact that conversion to the RGB format decomposes the spectrogram array into three amplitude components, which are then passed into the three parallel processing channels of the network. This way, complimentary information is analyzed by each channel rather than a copy of the same information given to each channel when raw spectrogram arrays (or greyscale images) are processed. To normalize the color intensity range across all images, the minimum and maximum values of the spectral magnitude were estimated over the entire speech database and mapped into the normalized dynamic range of the RGB images spanning from Min [dB] to Max [dB] values. The normalization step was critical in achieving good visualization of speech spectral components. The Min and Max values were chosen to maximize the subjective visibility of contours outlining time-frequency evolution of fundamental frequency (F0) speech formants and harmonic components of F0.

Since the required input size for ResNet-18 was 224 x 224 pixels, the original image arrays of 257 x 259 pixels were resized. Consistent with [23], [25], [26], the resizing was very small, causing no significant distortion to the spectrogram images and having no effect on the model training results. Each color component of the RGB images was passed as an input to a separate channel of ResNet-18.

#### 5) CNN MODEL

The classification was achieved using a pre-trained ResNet-18 CNN model, which has a residual network block architecture comprised of 18 layers. The network is defined by 11.5 million parameters; it requires a storage space of 44 MB. Natural image classification experiments based on ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) dataset [28], containing 1.2 million images with 1000 classes, demonstrated that the ResNet-18 model offers a very good balance between the computational cost and the quality of performance [29]. In comparison with other pre-trained CNNs, ResNet-18 achieves higher accuracy than popular models with linear architectures such as AlexNet [30]. It is also significantly faster than other residual block networks (ResNet-50, ResNet-101) without compromising the classification accuracy [24]. ResNet18 has shown good performance in comparison with other pre-trained CNN models in speech classification tasks such as, for example, intoxication detection [31], depression detection [32], and speech command recognition [33]. In the current study, the original last fully connected layer of ResNet-18, the Softmax layer, and the output layer were modified to have the number of class outputs required by a given experiment.

#### 6) TRAINING AND TESTING THE CNN MODEL

The entire dataset of speech samples was divided into the training/validation subset consisting of 80% of data and the testing sub-set composed of 20% of data. In both subsets, all speakers were represented in a balanced way. The testing data was not used in the training process. Each of the classification experiments described in Section IV was repeated three times with different mutually exclusive training and testing sets, and the results were given as average values over all repetitions.

### C. CLASSIFICATION PERFORMANCE MEASURES

To determine the quality of the classification outcomes, the accuracy $A_{c_i}$, was estimated for each class $c_i$ ($i = 1, \ldots, N$) using (3) [27].

$$A_{c_i} = \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i} \tag{3}$$

where, $N$ denotes the number of classes, $tp_i$ and $tn_i$ are the numbers of true-positive and true-negative classification outcomes, respectively. Similarly, $fp_i$ and $fn_i$ denote the numbers of false-positive and false-negative classification outcomes,

respectively. In cases when the classification was based on unbalanced data, the $F$-Score was calculated as,

$$F_{c_i} = 2\frac{p_{c_i} \times r_{c_i}}{p_{c_i} + r_{c_i}} \tag{4}$$

where, $p_{c_i}$, denotes the precision parameter given as,

$$p_{c_i} = \frac{tp_i}{tp_i + fp_i} \tag{5}$$

whereas $r_{c_i}$ is the recall parameter given as,

$$r_{c_i} = \frac{tp_i}{tp_i + fn_i} \tag{6}$$

## IV. RESULTS AND DISCUSSION

### A. EXPERIMENTAL SCHEDULE

This section describes experiments validating the concept of speech-based prediction of interpersonal trust and familiarity between team members. The experimental scenarios are explained in the following sub-sections.

### B. PREDICTION OF PRE- AND POST-MISSION TRUST USING SINGLE TRANSFER LEARNING

In this experiment, we wanted to determine if inter-personal trust can be predicted directly from the acoustic characteristics of speech. Speech recordings were automatically classified into two categories, Low/High trust, using a pre-trained CNN model ResNet-18. The ResNet model was pre-trained on over one million images to perform a general image classification task of object recognition. Here, we have fine-tuned this model to solve the more specific and more abstract task of trust prediction from our much smaller collection of spectrogram images. Due to the pre-existing relevant image classification knowledge embedded into the model, it was possible to accomplish the training (fine-tuning) process within a relatively short time and with a much smaller number of training images compared to what would be required to train the CNN model "from scratch" (i.e., without any prior knowledge built into the network). We refer to this approach as "single transfer learning." The training and classification tasks were performed in four separate cases corresponding to RIS, and TTP trust construct labels made before and after the mission (i.e., PreRIS, PreTTP, PostRIS, and PostTTP). The numbers of spectrogram images per class used to train the model were: 27661 for PreTTP, 27654 for PostTTP, 20935 for PreRIS, and 39087 for PostRIS. Fig. 3 shows the results of this experiment. General observations stemming from this figure are:

- In all cases of single transfer learning presented in Fig. 3, the trust classification accuracy is above 80%, varying from 82% to 86%. It indicates that acoustic speech characteristics are correlated with interpersonal trust. Therefore, speech classification can provide quite an efficient prediction of trust.
- Prediction of trust before the mission is only slightly lower than after the mission.

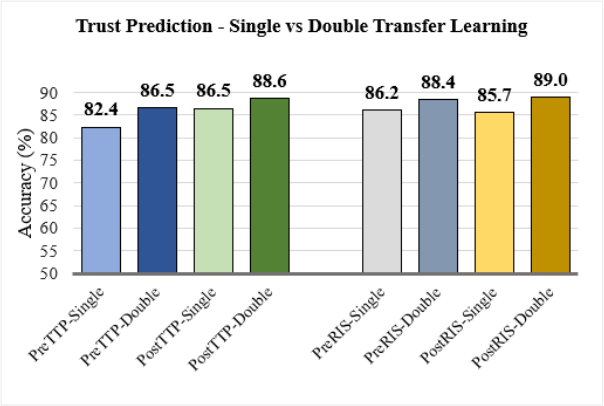


**FIGURE 3.** Prediction of trust between team members, using a single or double transfer learning. The double transfer learning was achieved by pre-training the network to recognize team familiarity.

- Before the mission, the RIS based prediction is 3.8% higher than the TTP based prediction. However, after the mission, both constructs give very similar outcomes.
- The best performing single transfer learning trust prediction models are given by the CNN trained on the PostTTP trust, and the CNNs trained on the PreRIS or PostRIS construct; in both cases, the achieved accuracy was about 86%.

### C. PREDICTION OF TEAM FAMILIARITY USING SINGLE TRANSFER LEARNING

Our aim in this experiment was to determine if acoustic speech characteristics are indicative of team familiarity. As in the first experiment, a single transfer learning was applied to predict if team members knew each other before the mission. However, this time, the pre-trained ResNet-18 model was fine-tuned to perform binary classification of speech into two categories Known/Unknown. The number of spectrogram images per class used to train the model was 34956. Fig. 4 shows that after training, the model could predict team members' familiarity with an accuracy of 85.16%. It supports the proposition that acoustic speech characteristics are correlated with team members' familiarity.

### D. PREDICTION OF PRE- AND POST-MISSION TRUST USING DOUBLE TRANSFER LEARNING

In this experiment, the aim was to determine if the addition of closely related pre-requisite knowledge into the model can improve the prediction of trust. From the previous two experiments (Sections IV-B&C), we knew that both interpersonal trust and team members' familiarity influence speech acoustics. Therefore, if there was a dependency between trust and familiarity, a network model having a pre-requisite knowledge of team members' familiarity could hypothetically outperform a model that did not have such a pre-requisite knowledge. To investigate this assumption, we have conducted a double transfer learning experiment. The CNN model generated through transfer learning from ResNet-18

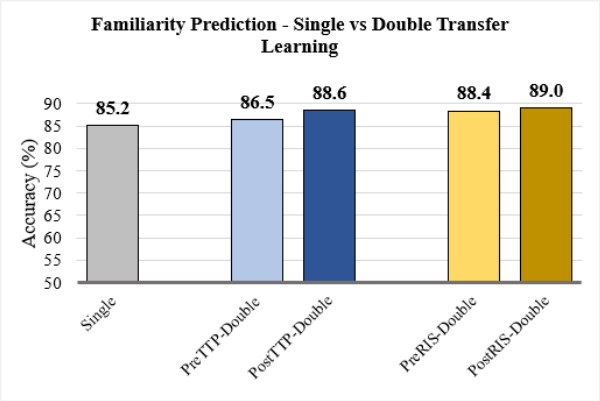


**FIGURE 4.** Prediction of team members familiarity, using a single or double transfer learning. The double transfer learning was achieved by pre-training the network to recognize TTP or RIS trust (assessed pre- or post-experiment).

(described in Section IV-C) to predict prior knowledge of team members was further trained to predict trust between team members. The transfer of knowledge during this process occurred twice, the first time when transferring the general image object classification knowledge from ResNet-18 to train the CNN model to recognize if the team members know each other, and the second time when this knowledge was transferred to further train the CNN model to recognize trust between team members. We refer to this approach as double transfer learning.

The training and classification tasks within this scenario were conducted in four cases. This included the prediction of RIS and TTP constructs labeled before and after the mission described by four types of binary trust/no-trust labels (PreTTp, PostTTP, PreRIS, and PostRIS). The numbers of spectrogram images per class used to train the model were: 27661 for PreTTP, 27654 for PostTTP, 20935 for PreRIS, and 39087 for PostRIS. The results are presented in Fig. 3, give the following general observations:

- In all four cases, the double transfer learning provided improvement upon the single transfer learning approach. For PreTTP trust the improvement was the highest, by 4.1%, for PostTTP by 2.1%, for PreRIS by 2.2%, and for PostRIS by 3.3%. It indicates that there is a strong correlation between interpersonal trust and pre-requisite familiarity between team members. Therefore, a network model having knowledge of familiarity provided a better prediction of trust than a network predicting without this knowledge.
- Like in the case of single transfer learning, the double learning prediction of trust before the mission appears to be slightly lower than after the mission.
- Prediction of trust before the mission when using the TTP construct is 1.9% higher when using the RIS than when using the TTP construct. This difference is smaller compared to the single transfer learning (Section IV-B). After the mission, both constructs TTP and RIS give similarly high outcomes of about 89%.
- The best performing double transfer learning trust prediction model was given by the CNN fine-tuned on the PostRIS trust construct; it achieved an accuracy of 89%.

### E. PREDICTION OF TEAM FAMILIARITY USING DOUBLE TRANSFER LEARNING

This experiment was a reversed version of the experiment described in Section IV-D. Our aim was to find out if the addition of related pre-requisite knowledge of interpersonal trust can improve prediction of the pre-mission familiarity between team members. The CNN model generated through transfer learning from ResNet-18 (described in Section IV-B) to predict trust between team members was further trained to predict the pre-missional familiarity of team members. This included the prediction of familiarity based on models pre-trained on four different trust constructs (PreTTp, PostTTP, PreRIS, and PostRIS). The number of spectrogram images per class used to train the model was 34956 for PreTTP, PostTTP, PreRIS, and PostRIS constructs. The results are presented in Fig. 4 lead to the following observations:

- Compared to the single transfer learning baseline accuracy of 85.2%, in all four cases, double transfer learning increased the prediction accuracy of team members' familiarity. For the model pre-trained on the PreTTP trust construct, the increase was by 1.3% for the PostTTP by 3.4%, for the PreRIS by 3.2%, and for the PostRIS by 3.8%. It indicates that pre-requisite knowledge of interpersonal trust can improve the prediction of the pre-mission familiarity between team members.
- The best performing double transfer learning familiarity prediction model was given by the CNN fine-tuned on the PostRIS trust construct; it achieved an accuracy of 89%.

### F. JOINT PREDICTION OF TRUST AND TEAM FAMILIARITY USING SINGLE TRANSFER LEARNING

To observe the difference between the concept of double transfer learning between trust and familiarity, and a simultaneous prediction of these two categories, we have trained a separate CNN model to perform a joint prediction of trust and familiarity based on a single transfer learning. Unlike in double transfer learning, where there was a gradual build-up of cognition components, in the multilabel classification, all components were generated simultaneously. The speech labels and the number of spectrogram images for each class used in the process of training are shown in Table 4. The results in Fig. 5 show the following:

- The achieved accuracy was ranging between 80% and 84% depending on the type of the applied trust construct. Given that for the four classes, the pure guess was 25%, these results are relatively high. It confirms that both trust and familiarity are affecting speech acoustics.
- The RIS based prediction was higher than the TTP based prediction (by 5% for the PreRIS and by 1.2% for the PostRIS).

**TABLE 4.** Numbers of Training Images Across Classes in Joint Prediction of Trust and Team Members Familiarity.

| Class (Trust/Familiarity) | PreTTP | PreRIS | PostTTP | PostRIS |
|---|---|---|---|---|
| Low-Known | 25468 | 20935 | 32823 | 31160 |
| Low-Unknown | 2193 | 0 | 20002 | 7927 |
| High-Known | 19965 | 24498 | 12610 | 14273 |
| High-Unknown | 32763 | 34956 | 14954 | 27029 |

**TABLE 5.** Numbers of Participants Who Changed or Not Their Trust in Partner Assessment After the Experiment.

| Trust Change | TTP | | RIS | |
|---|---|---|---|---|
| | Known | Unknown | Known | Unknown |
| No Change | 14 | 14 | 15 | 21 |
| High to Low | 8 | 14 | 9 | 7 |
| Low to High | 4 | 0 | 2 | 0 |

### G. PREDICTION OF CHANGE IN TRUST

One of the observations given by the data characteristics in Fig. 1 is that some of the participants changed their trust in partners after the mission. As shown in Table 5, 28 participants (51.8% of the total of 57 participants) did not change their TTP trust assessment, and 36 (66.6%) did not change their RIS trust. The shift from High to Low trust occurred in 24 cases (44.4%) of TTP trust and in 16 cases (29.6%) of RIS trust. The opposite shift from Low to High trust was observed in 4 TTP cases (7.02%) and in 2 RIS cases (7.4%). Generally, the majority of participants did not change their trust assessment. When the change occurred, it was mostly from High to Low and only in very few cases from Low to High. We have already shown in Sections IV-B&D that speech classification can be used to predict trust level; however, in this experiment, we wanted to find out if we can use speech classification to predict a change in the trust after the problem-solving mission. Speech data was, in this case, labeled as Change/No-Change. 38744 images per class were used to train the model based on TTP trust construct, and 26624 images for the model based on RIS construct. The results illustrated in Fig. 6 show that the trust change prediction based on the RIS construct achieved 86% accuracy; thus, outperforming the TTP based prediction by 3.59%. In both cases, the prediction accuracy was above 80%, confirming that speech classification can be used to obtain a good prediction of trust change.

### H. JOINT PREDICTION OF TRUST, TEAM FAMILIARITY, AND TRUST CHANGE

In this experiment, we tested a more complex classification scenario of simultaneous prediction of trust, familiarity, and trust change given eight different class labels, as listed in Table 6. Due to the fact that there was no data available to represent some of the classes, the classification was

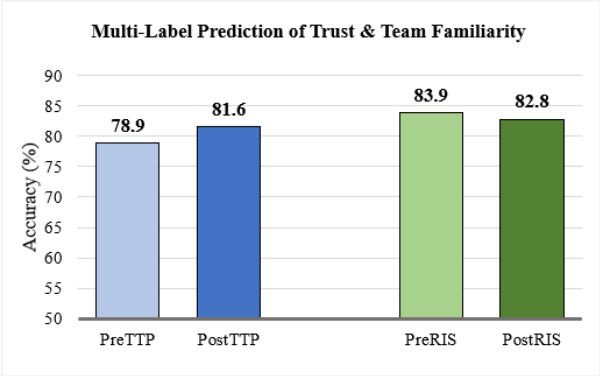


**FIGURE 5.** Multi-label prediction of trust (PreTTP and PostRIS) and team familiarity.

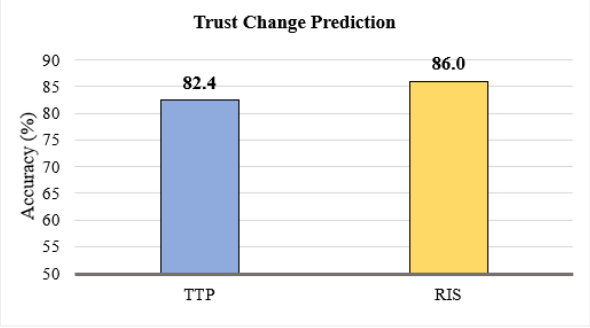


**FIGURE 6.** Prediction of trust (TTP and RIS) change before and after experiment.

**TABLE 6.** Numbers of Training Images Across Classes in Joint Prediction of Trust and Team Members Familiarity and Trust Change.

| Class (Trust/Familiarity/ Change) | Pre TTP | Pre RIS | Post TTP | Post RIS |
|---|---|---|---|---|
| Low-Known-Change | 6790 | 4236 | 14145 | 14461 |
| Low-Known-No Change | 18678 | 16699 | 18678 | 16699 |
| Low-Unknown-Change | 0 | 0 | 17809 | 7927 |
| Low-Unknown-No Change | 2193 | 0 | 2193 | 0 |
| High-Known-Change | 14145 | 14461 | 6790 | 4236 |
| High-Known-No Change | 5820 | 10037 | 5820 | 10037 |
| High-Unknown-Change | 17809 | 7927 | 14954 | 27029 |
| High-Unknown-No change | 14954 | 27029 | 0 | 0 |

unbalanced; therefore, the results of this experiment shown in Fig. 7 include both accuracy and F-scores. Both accuracy and F-score values shown in Fig. 7 are within a narrow range of 80%-82%, indicating that in all four cases (PreTTP, PostTTP, PreRIS, and PostRIS), the classification outcomes were very similar. Given that the accuracy and F-score values were very close to each other, the imbalanced cases did not affect the overall performance in a significant way, and there was a good balance between false positive and false negative classification outcomes. As expected, the increased number

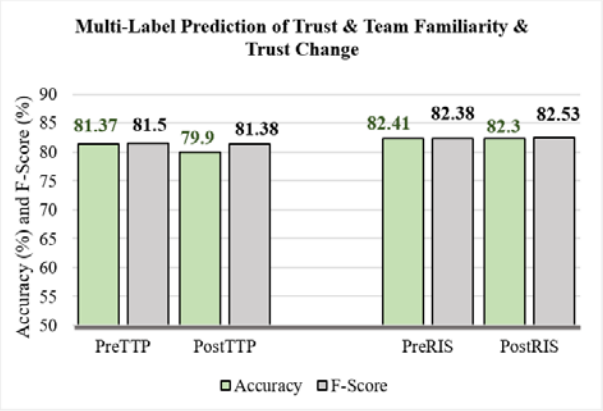


**FIGURE 7.** Multi-label prediction of trust (pre and post TTP and RIS), team familiarity and trust change.

of classes led to a slightly reduced classification accuracy (by about 1%) compared to two-class prediction (trust and familiarity) in Fig. 5.

## V. CONCLUSION

We have investigated acoustic speech characteristics as indicators of interpersonal trust and familiarity. A number of CNN models were trained to differentiate between different classes of trust, team members' familiarity with each other, and trust change after a team-based problem-solving mission.

The experimental results provided a strong indication that all three categories can be efficiently predicted from speech. In addition, we have shown that models having a pre-requisite knowledge of team members' familiarity can be more efficiently trained to predict trust between team members compared with models not having such pre-requisite knowledge. Similarly, models having a pre-requisite knowledge of trust between team members can be more efficiently trained to predict familiarity.

In general terms, the results of our study indicate that the more related pre-requisite knowledge that is embedded in the model, the faster and less data-consuming is the learning of a new task. Using an analogy with human learning, for example, a person who already knows three different languages is likely to learn a new language faster and with fewer examples than a person who knows one language only. Likewise, a tennis player can most likely learn baseball faster than a person who never played any sports before. In both examples, the prior, task-related knowledge is the key factor increasing the learning efficiency.

One of the important general implications of our observations is that the embedding of pre-requisite, related knowledge into the model can be viewed as a way of dealing with the scarcity of specific task-related training data. Namely, instead of increasing the training set size by generating synthetic or augmented data, the model can be enriched with a pre-requisite knowledge and learn from the small in size, but actual data.

Our future research will investigate applications of the findings described here into human-machine interactions. We will also look into potential applications of multiple transfer learning to counteracting the data imbalance in the automatic monitoring of human-machine interactions.

## ACKNOWLEDGMENT

The data used in this study was collected at the AFRL, Wright Patterson Air Force Base, OH, USA. The authors would like to thank Dr. Joe Lyons, AFRL, for his valuable comments and all the AFRL staff who facilitated the data collection and transfer to Australia. In addition, the authors would like to thank Dr. Benjamin Knott, AFOSR, and Dr. Mark Draper, AFRL, for their contributions to establishing an international research collaboration between AFOSR and DST.

## REFERENCES

[1] R. Borum, *The Science of Interpersonal Trust*, vol. 574. New York, NY, USA: Mental Health Law & Policy Faculty Publications, 2010. [Online]. Available: https://scholarcommons.usf.edu/mhlp_facpub/574

[2] A. Mislin, L. V. Williams, and B. A. Shaughnessy, "Motivating trust: Can mood and incentives increase interpersonal trust?" *J. Behav. Exp. Econ.*, vol. 58, pp. 11–19, Oct. 2015.

[3] N. Ashkanasy, "Trust and distrust in organizations: Trust and distrust in organizations: Dilemmas and approaches," *Personnel Psychol.*, vol. 58, no. 2, pp. 521–526, 2005, doi: 10.1111/j.1744-6570.2005.20050504_2.x.

[4] M. S. S. Syed, M. Stolar, E. Pirogova, and M. Lech, "Speech acoustic features characterising individuals with high and low public trust," in *Proc. 13th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, Gold Coast, QLD, Australia, Dec. 2019, pp. 1–9, doi: 10.1109/ICSPCS47537.2019.9008747.

[5] J.-H. Cho, K. Chan, and S. Adali, "A survey on trust modeling," *ACM Comput. Surv.*, vol. 48, no. 2, pp. 1–40, Nov. 2015, doi: 10.1145/2815595.

[6] A. Hussein, S. Elsawah, and H. Abbass, "Towards trust-aware human-automation interaction: An overview of the potential of computational trust models," in *Proc. 53rd Hawaii Int. Conf. Syst. Sci.*, 2020, pp. 1–10, doi: 10.24251/HICSS.2020.047.

[7] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity, and social history," *Games Econ. Behav.*, vol. 10, no. 1, pp. 122–142, Jul. 1995.

[8] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *Acad. Manage. Rev.*, vol. 20, no. 3, pp. 709–734, Jul. 1995, doi: 10.5465/amr.1995.9508080335.

[9] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships," *J. Personality Social Psychol.*, vol. 49, no. 1, pp. 95–112, 1985.

[10] L. M. DeBruine, "Trustworthy but not lust-worthy: Context-specific effects of facial resemblance," *Proc. Roy. Soc. B, Biol. Sci.*, vol. 272, no. 1566, pp. 919–922, 2005, doi: 10.1098/rspb.2004.3003.

[11] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues," in *Proc. Interspeech*, Sep. 2018, pp. 416–420.

[12] P. Belin, B. Boehme, and P. McAleer, "The sound of trustworthiness: Acoustic-based modulation of perceived voice personality," *PLoS ONE*, vol. 12, no. 10, Oct. 2017, Art. no. e0185651, doi: 10.1371/journal.pone.0185651.

[13] A. C. Elkins and D. C. Derrick, "The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents," *Group Decis. Negotiation*, vol. 22, no. 5, pp. 897–913, Sep. 2013, doi: 10.1007/s10726-012-9339-x.

[14] A. Schirmer, M. H. Chiu, C. Lo, Y.-J. Feng, and T. B. Penney, "Angry, old, male—And trustworthy? How expressive and person voice characteristics shape listener trust," *PLoS ONE*, vol. 15, no. 5, May 2020, Art. no. e0232431, doi: 10.1371/journal.pone.0232431.

[15] X. Jiang, K. Gossack-Keenan, and M. D. Pell, "To believe or not to believe? How voice and accent information in speech alter listener impressions of trust," *Quart. J. Exp. Psychol.*, vol. 73, no. 1, pp. 55–79, Jan. 2020, doi: 10.1177/1747021819865833.

[16] B. Waber, M. Williams, and J. S. Carroll, "A voice is worth a thousand words: The implications of the micro-coding of social signals in speech for trust research," in *Handbook of Research Methods on Trust*. Cheltenham, U.K.: Edward Elgar Publishing, 2015, doi: 10.4337/9781782547419.00037.

[17] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. Interspeech, 8th Annu. Conf. Int. Speech Commun. Assoc.* Antwerp, Belgium: ISCA, 2007, pp. 2253–2256.
[18] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers Psychol.*, vol. 4, pp. 1–12, May 2013.
[19] J. J. Lee, W. B. Knox, J. B. Wormwood, C. Breazeal, and D. DeSteno, "Computationally modeling interpersonal trust," *Frontiers Psychol.*, vol. 4, p. 893, Dec. 2013, doi: 10.3389/fpsyg.2013.00893.
[20] H. M. Khalid, W. S. Liew, M. G. Helander, and C. K. Loo, "Prediction of trust in scripted dialogs using neuro-fuzzy method," in *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manage. (IEEM)*, Dec. 2016, pp. 1558–1562, doi: 10.1109/IEEM.2016.7798139.
[21] H. M. Khalid, L. W. Shiung, P. Nooralishahi, Z. Rasool, M. G. Helander, L. C. Kiong, and C. Ai-Vyrn, "Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 60, no. 1. Los Angeles, CA, USA: Sage, 2016, pp. 697–701, doi: 10.1177/1541931213601160.
[22] A. R. Panganiban, G. Matthews, and M. D. Long, "Transparency in autonomous teammates: Intention to support as teaming information," *J. Cognit. Eng. Decis. Making*, vol. 14, no. 2, pp. 174–190, Jun. 2020.
[23] M. Lech, M. Stolar, R. Bolia, and M. Skinner, "Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 3, no. 4, pp. 363–371, 2018, doi: 10.25046/aj030437.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[25] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding," *Frontiers Comput. Sci.*, vol. 2, pp. 1–14, May 2020, doi: 10.3389/fcomp.2020.00014.
[26] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, Aug. 2017.
[27] B. Schuller, S. Steidl, and A. A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.
[28] 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," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
[29] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," 2016, *arXiv:1605.07678*. [Online]. Available: http://arxiv.org/abs/1605.07678
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105, doi: 10.1061/(ASCE)GT.1943-5606.0001284.
[31] W. Wang, H. Wu, and M. Li, "Deep neural networks with batch speaker normalization for intoxicated speech detection," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Lanzhou, China, Nov. 2019, pp. 1323–1327, doi: 10.1109/APSIPAASC47483.2019.9023074.
[32] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Comput. Sci.*, vol. 164, pp. 618–628, 2019.
[33] B. McMahan and D. Rao, "Listening to the world improves speech command recognition," Dept. Comput. Sci., Sound, Cornell Univ., Ithaca, NY, USA, Tech. Rep., 2017, pp. 1–8. [Online]. Available: https://arxiv.org/abs/1710.08377

**CATHERINE SANDOVAL RODRIGUEZ** (Graduate Student Member, IEEE) received the B.S. degree in electronic engineering from Pontificia Universidad Javeriana, Colombia, in 2001, and the M.S. degree in electronic engineering from RMIT University, Australia, in 2015, where she is currently pursuing the Ph.D. degree in electrical and electronic engineering. Her research interests include artificial intelligence, machine learning, fine art analysis, deep learning, and image processing.

**APRIL ROSE PANGANIBAN** received the B.S. degree in psychology from the University of Florida, in 2002, and the M.A. and Ph.D. degrees in experimental and human factors psychology from the University of Cincinnati. Since 2007, she has been working with the Air Force Research Laboratory, as a Contractor, where she has also been a Civilian Research Psychologist, since 2013. Her research interests include individual differences, specifically, executive functions and stress, as well as team communication, in both human–human and human–machine groups.

**MELISSA N. STOLAR** received the B.E. degree in science and electronic engineering and the Ph.D. degree in electrical engineering from the Royal Melbourne Institute of Technology University (RMIT), Melbourne, Australia, in 2013 and 2017, respectively. From 2016 to 2019, she was a Research Fellow with the School of Engineering, RMIT University. Since 2020, she has been a Data Scientist with the Defence Science Technology Group, Melbourne. Her research interests include audio-visual signal processing, neural networks, and machine learning.

**ROBERT BOLIA** received the B.A. degree in mathematics from Wright State University, in 1997, and the M.A. degree in military studies (Joint Warfare) from American Military University, in 2004. From 1989 to 2008, he was a Research Scientist with the Air Force Research Laboratory, Human Effectiveness Directorate, USA. From 2008 to 2016, he served as the Associate Director of the Office of Naval Research Global, Tokyo, Japan, and Santiago, Chile. In 2016, he joined the Defence Science and Technology Group, Melbourne, Australia. He served as a Group Leader of Human Factors, until 2018, and is currently a Research Leader of Aerospace Decision Effectiveness. He also leads the DST Strategic Research Initiative on Trusted Autonomous Systems. His research interests include cognitive science, decision making, linguistics, and military history.

**MARGARET LECH** (Member, IEEE) received the M.S. degree in physics from Maria Curie-Skłodowska University, Poland, and the Ph.D. degree in electrical engineering from The University of Melbourne, Australia. She is currently a Professor with the School of Engineering, RMIT University, Australia. Her research interests include machine learning applications in speech and image processing, system modeling, and optimization.