AFRL-AFOSR-JP-TR-2018-0062

Fostering positive team behaviors in human-machine teams through emotion processing: Adapting to the operator's state

Margaret Lech
ROYAL MELBOURNE INSTITUTE OF TECHNOLOGY
124 LATROBE ST
MELBOURNE, 3000
AU

08/17/2018
Final Report

Air Force Research Laboratory
Air Force Office of Scientific Research

Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
17-08-2018	Final	22 Feb 2017 to 21 Mar 2019

4. TITLE AND SUBTITLE
Fostering positive team behaviors in human-machine teams through emotion processing: Adapting to the operator's state

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-17-1-0095

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Margaret Lech, April Rose Fallon

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
ROYAL MELBOURNE INSTITUTE OF TECHNOLOGY
124 LATROBE ST
MELBOURNE, 3000 AU

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2018-0062

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The team developed a software system that simultaneously recognizes 7 emotional categories as speech is produced. It is suitable for applications on cellular phones and online speech communication platforms. The methodology uses deep learning (DL) with speech signals being represented in the form of RGB images of speech spectrograms. By representing speech signals in the form of RGB images, the speech classification problem was re-defined as an image classification task. This created an opportunity to replace the lengthy and data-costly training of a deep neural network by the shortened and more data-efficient fine tuning of an existing pre-trained image classification network (AlexNet). The speech emotion recognition (SER) results achieved with the fine-tuned AlexNet (FTAlexNet) showed an average accuracy of 80% for the Berlin Emotional Speech data. This result was found to be comparable with existing state-of-the art techniques, but with the advantage of significantly lower computational and data costs.

The ability to analyze emotions represented in speech was then applied to a multi-stage classification system with intermediate learning (MSIL). In this scheme, the system can leverage the mistakes made by the primary stage in learning from the data and use them to improve the learning by the secondary stage. It is felt that this approach can be incorporated as a building block for more complex multi-level machine reasoning systems.

15. SUBJECT TERMS
Multimodal, Human-Machine Team, Speech Perception

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			RIECKEN, RICHARD
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 703-941-1100

Type of Report	Final Performance Report
Project Title:	Fostering positive team behaviors in human-machine teams through emotion processing: Adapting to the operator's state
Project ID:	AOARD Project Code: FA2386-17-1-0095
Reporting Period:	From 22 February 2017 to 19 June 2018

Fostering positive team behaviors in human-machine teams through emotion processing: Adapting to the operator's state

Background:

Machines' inability to understand non-linguistic information in people's voices and facial images reduces human trust in machines.

Objective(s):

- Design algorithms capable of detecting and predicting human emotions from audio-visual data labeled with categorical emotions.
- Determine and model links between a person's emotional and physiological characteristics, and trustworthiness.
- Design algorithms predicting trust from audio-visual (speech and image) and biometric (EEG, ECG, GSR) data.
- Determine which predictors of trust derived from human-human data can be used to foster positive team behaviors in human-machine teams.

Impact to AF:

- Objective measures and models of inter-personal behavior that can be applied to increase trust in machine autonomy.

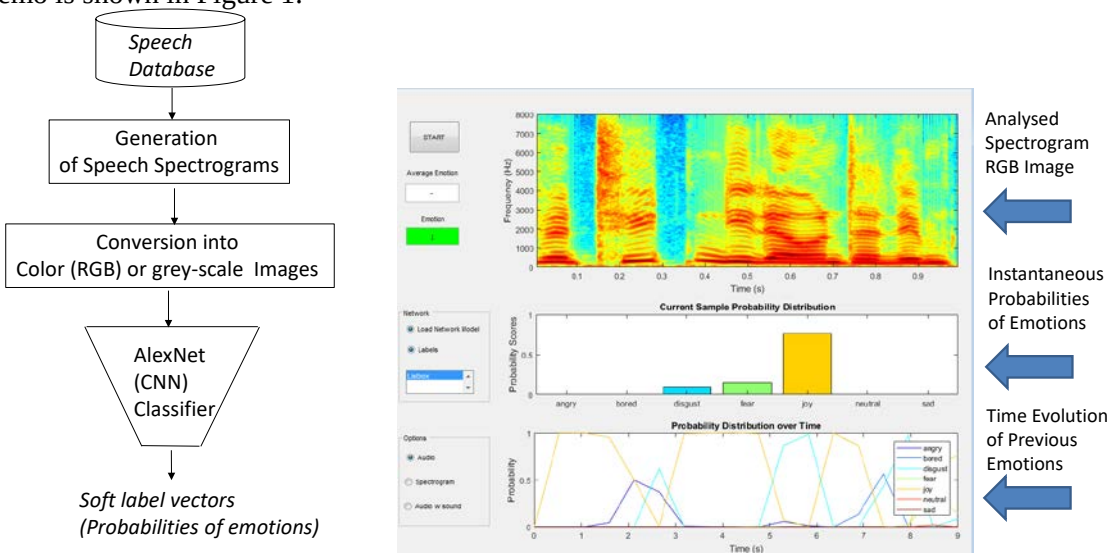
AFOSR-DSTG co-operation highlights

- 14-17 March 2016, M. Lech gave two research presentations at the AFRL, Dayton, Ohio, U.S.A.
- February 2017, AOARD provided USD 25,000 to support the collaboration from February 2017 to February 2018.
- April 2017-current, regular meetings through Skype have been conducted between M. Lech (RMIT), M. Stolar (RMIT) and A. R. Panganiban (AFRL). Details of software design and experimental protocol for the Dyadic Team Trust (DDT) data collection were discussed. The meetings will be continued in 2018-2019.
- 15 February 2018, AOARD approved further USD 25,000 to support the collaboration from February 2018 to February 2019.
- 30 February 2018, School of Engineering, RMIT allowed for up to 3 PhD scholarships to support the RMIT- AFOSR – DSTG research cooperation. One PhD candidate has already commenced and two potential candidates are under review.
- 5 May 2018, software design for the DDT was completed and pilot experiments provided preliminary data samples.
- 30 April – 5 May, 2018, A.R. Panganiban visited RMIT University, Melbourne, Australia to discuss results of the PPT pilot data and plans for future collaboration 2018-2019.
- 3-4 May 2018, A.R Panganiban, M. Lech, M. Stolar, R. Bolia and M. Skinner participated in the 1st Annual Review and Workshop AFOSR – DSTG Co-Sponsored Research Program on Trusted Autonomy 3-4 May 2018, Royal Melbourne Institute of Technology, Melbourne, Victoria, Australia.

- In August 2018, M. Stolar is expected to have a 3-weeks research visit to AFRL, Dayton, Ohio, U.S.A. to work with A.R. Panganiban. The visit will be supported by the Windows-on-Science program (AFOSR).

Summary of Research Activities:

1. Developed software for a low-cost, real-time, speech emotion recognition system; published in [1]. The system simultaneously recognises 7 emotional categories as speech is produced. It is suitable for applications on cellular phones and online speech communication platforms. The methodology uses deep learning (DL) with speech signals being represented in the form of RGB images of speech spectrograms. By representing speech signals in the form of RGB images, the speech classification problem was re-defined as an image classification task. This created an opportunity to replace the lengthy and data-costly training of a deep neural network by the shortened and more data-efficient fine tuning of an existing pre-trained image classification network (AlexNet). The SER results achieved with the fine-tuned AlexNet (FTAlexNet) showed an average accuracy of 80% for the Berlin Emotional Speech data [1]. This result was found to be comparable with existing state-of-the art techniques, but with the advantage of significantly lower computational and data costs. A demo of the real-time application of the FTAlexNet can be viewed on: (https://www.youtube.com/watch?v=fuMpF3cUqDU&t=6s). An example screen-shot of this demo is shown in Figure 1.



(a) (b)

Figure 1. Real-time speech emotion recognition. Block diagram (a). An example screen shot of a real-time application of the FTAlexNet (b). The top window shows the currently analysed image, the middle window shows estimates of the instantaneous probabilities of 7 emotional states for the analysed image (dominated by joy in this example), and the bottom window shows the time history of past emotion probabilities.

2. Completed analysis of amplitude-frequency characteristics of emotional speech; under review in [3], [4]. Automatic speech emotion recognition (SER) techniques based on acoustic analysis show high confusion between certain emotional categories. This study used an indirect approach to determine the differences between the amplitude-frequency characteristics of different emotions in order to support the development of future, more efficient, SER methods. The analysis was carried out by representing short-time frames of speech in the form of color RGB images of speech spectrograms, which were used to fine-tune a pre-trained image classification network to recognize emotions. Spectrogram representation on four different frequency scales - linear, melodic, equivalent rectangular bandwidth (ERB) and logarithmic - allowed observation of the effects of more- or less-detailed visualization of low- or high-frequency characteristics of speech. Use of either the red (R), green (G) or blue (B) components of the RGB images showed the importance of different amplitude levels. Experiments conducted on the Berlin emotional speech data (EMO-DB) revealed the most likely locations of seven emotional categories (anger, boredom, disgust, fear, joy, and neutral) on the amplitude-frequency plane.

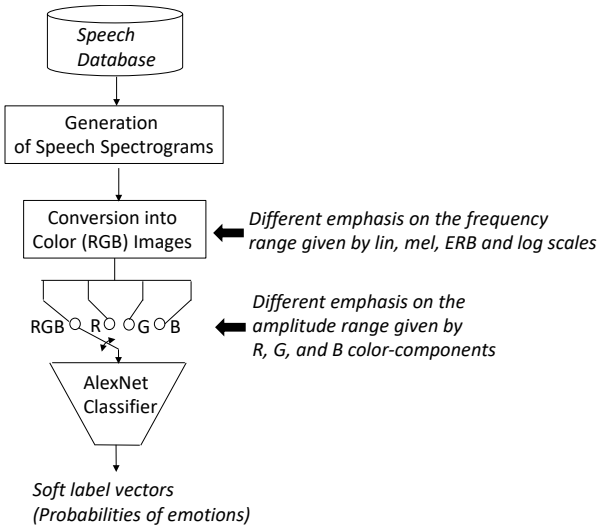


Figure 2. Block diagram of the amplitude-frequency analysis system using classification of images of speech spectrograms.

3. Tested a new concept of a multi-stage classification system with intermediate learning (MSIL); published in [2]. The aim of this study was to take a step towards creating an autonomous machine reasoning system by designing a simple multiple-stage classification method with intermediate learning (MSIL). While the first stage performs learning based on primary feature parameters derived from the data, the second stage performs learning based on classification labels given by the first stage (see Figure 2). This means that the second stage learns how to compensate for the inter-class confusion patterns occurring at the first stage. This is an important and useful property, since in many classification problems it is not clear what features provide optimal performance. In such cases, the MSIL approach could be beneficial by compensating for the inadequacies of feature parameters used at the first stage of learning. A two-stage version of the MSIL was tested on nine datasets commonly used in the validation of machine learning tasks. In most cases, the results showed an incremental improvement of classification results between subsequent stages of the classification unit. This means that the primary learning from the data can be improved by the secondary learning from mistakes made when classifying the data parameters. The MSILs could provide a basis for creating more complex structures of algebraically connected MSIL units, where the flow of data between these units would infer the problem-solving outputs. Given the general character of the proposed unit, it can be used as a building block for more complex multi-level and multi-modal machine reasoning systems.

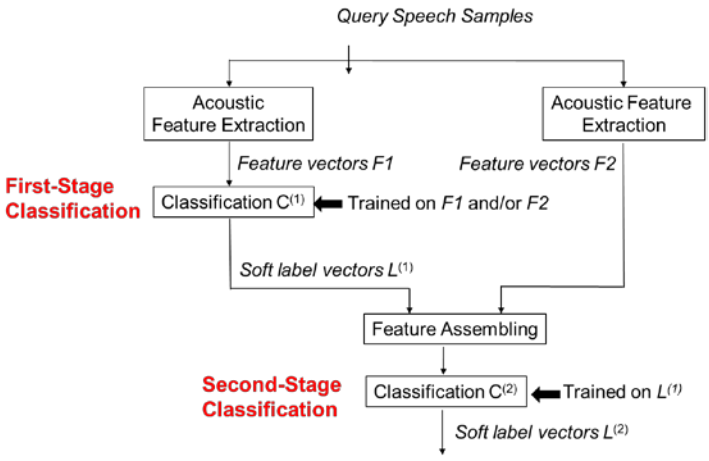


Figure 3. Multi-stage classification with intermediate learning (MSIL).

4. Dyadic team trust data (DTT) collection by AF.
Dr April Rose Fallon (USAF AFMC 711 HPW/RHXS) designed an experiment to collect audio-visual data labelled with emotional and trust states. The experiment will collect data from human-human teams playing a computer game and making joint decisions requiring a certain level of trust between team members. During the decision-making stage, team members will report their emotions and trust levels. In addition, physiological information including EEG, ECG, GSR and facial video images will be collected during the game. The AFOSR software engineering team has completed the software development for this experiment. The data collection is in progress. Dr Fallon will visit RMIT University at the end of April 2018 for about 1 week. She will discuss with our team how the AF data can be analyzed to predict trust between team members, and how to determine potential links between trust and other subjective and physiological measures collected during her experiment.

5. Public trust in politicians (PTP) data collection from public sources.
We have made a substantial progress towards collecting data from politicians. The data includes audio-visual recordings of public speeches, interviews, transcripts of speeches and twitter texts collected from 30 politicians (mostly US Senators). Each politician has been ranked by at least 100 people via an online ranker (https://www.ranker.com/list/trustworthy-politicians/mike-rothschild). The group includes 10 high-trust, 10 mid-trust and 10 low-trust politicians. This data will allow us to determine if public perception of trustworthiness can be predicted acoustic and/or linguistic characteristics of a person's speech and facial expressions.

6. Software development for a multi-task learning (MTL) system.
We have made substantial progress towards software development that will be used to predict trust in a multi-tasking approach, where recognition of trust is one of many tasks to learn and solve simultaneously by a single MTL system. For example, the assessment of trustworthiness can be conducted in parallel with the recognition of gender, person, language, emotion, EEG, ECG and GSR. Simultaneous learning has been shown to outperform single-task systems. In addition, small and large datasets from various sources and with different labels can be combined to enhance the learning process. The development is based on the multi-task learning via structural regularization (MALSAR) programming package (http://www.yelab.net/software/MALSAR/).

Publications with acknowledgements of funding from AOARD

Published

[1]. MN Stolar, M Lech, RS Bolia and M Skinner, "Real Time Speech Emotion Recognition Using RGB Image Classification and Transfer Learning", Proceedings of the 11th International Conference on Signal Processing and Communication Systems, ICSPCS'2017, 13-15 December, Surfers Paradise, Australia, pp. 1-8.

[2]. MN Stolar, M Lech, RS Bolia and M Skinner, "Towards Autonomous Machine Reasoning: Multi-Stage Classification System with Intermediate Learning", Proceedings of the 11th International Conference on Signal Processing and Communication Systems, ICSPCS'2017, 13-15 December, Surfers Paradise, Australia, pp. 1-6.

Under Review

[3]. MN Stolar, M Lech, RS Bolia and M Skinner, "Emotion Recognition from a Streaming Speech Signal Using a Pre-Trained Image Classification Network". Journal of Acoustical Society of America (JASA).

[4]. MN Stolar, M Lech, RS Bolia and M Skinner, "Amplitude-Frequency Characteristics of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images", Advances in Science, Technology and Engineering Systems (ASTES) Journal.

Outlines of Future Research (from 22 March 2018 to 21 March 2019):

1. Finalise data collection for both "trust" data sets DTT and PTT.
2. Determine acoustic, linguistic and physiological correlates of trust.
3. Design algorithms that learn to predict trust from multi-modal (speech, text, facial images and physiological information) data.
4. Determine which predictors of trust can be used to foster positive team behaviours in human-machine teams.