IberSPEECH 2022
14-16 November 2022, Granada, Spain

Measuring trust at zero-acquaintance using acted-emotional videos

Cristina Luna-Jiménez[1], *Syaheerah Lebai Lutfi*[2], *Manuel Gil-Martín*[1], *Ricardo Kleinlein*[1], *Juan M. Montero*[1], *and Fernando Fernández-Martínez*[1]

[1] Grupo de Tecnología del Habla y Aprendizaje Automático (T.H.A.U. Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid, 28040, Madrid, Spain; [2]School of Computer Sciences, Universiti Sains Malaysia, 11800, Minden, Pulau Pinang, Malaysia

cristina.lunaj@upm.es, syaheerah@usm.my, manuel.gilmartin@upm.es, ricardo.kleinlein@upm.es, juanmanuel.montero@upm.es and fernando.fernandezm@upm.es

Abstract

Trustworthiness recognition attracts the attention of the research community due to its main role in social communications. However, few datasets are available and there are still many dimensions of trust to investigate. This paper presents a study of an annotation tool for creating of a trustworthiness corpus. Specifically, we asked the participants to rate short-emotional videos extracted from RAVDESS at zero acquaintance and studied the relationship between their trustworthiness score and other characteristics of the subjects of each video. Eloquence ($\rho = 0.41$), kindness ($\rho = 0.32$), attractiveness ($\rho = 0.34$), and authenticity of emotion transmitted ($\rho = 0.6$) are shown to be important determinants of perceived trustworthiness. In addition, we have measured a strong association between some of the variables under study. For example, physical beauty and voice pleasantness obtain a $\rho = 0.71$, or eloquence and expressiveness ($\rho = 0.65$), which opens a future line of investigation to study how people understand attractiveness and eloquence from these perspectives. Finally, an attribute selection strategy identified that frequency and spectral-related attributes could be accurate aural indicators of perceived trustworthiness.

Index Terms: trustworthiness, emotion, annotation platform, trust recognition, affective computing, HMI

1. Introduction

Trustworthiness is still an open issue for the research community in various domains, from psychology to computer sciences. However, trust is critical in our daily life and in negotiations [1, 2]. For this reason, the number of investigations studying this topic is currently increasing [3, 4].

Having perceived trust recognizers could be a valuable resource for its exploitation in diverse fields. For example, in the health care area, having Conversational Agents able to transmit trust could encourage patients to give more sensitive information, ensuring faster treatments and disease diagnosing [5]. Also, designing systems able to elicit trust or detect people's affective state in emergencies could save lives due to a reduction in action times [6]. A similar effect could be observed in human resources recruiters. Recruiters perceived as more trustworthy could obtain more sincere answers from the interviewees, facilitating the recruitment decision and the suitability of the person for the position.

Due to its many applications and the few resources available, this paper aims to present a preliminary study in the line of acquiring a dataset with perceived trustworthiness annotations at zero acquaintance. Section 3 explains the developed data acquisition platform to annotate perceived trustworthiness in short emotional videos at zero acquaintance (only based on first impressions without previous contact or interaction with the person on the video). Next, in Section 4, we analyze the results of the correlation matrices and the suitability of the tested surveys for the task. We also show an experiment on the aural features that most correlate with perceived trustworthiness. Finally, in Section 5 and Section 6 we summarize the limitations of the current work and drag the main conclusions.

2. Related Works

From the psychological perspective, there are many attempts to define trust based on descriptions or its characteristics, such as its subjectivity, its relationship character, its dynamism and context dependability, and its necessity of uncertainty or risk, from which we could obtain a benefit through a trustworthiness relationship; among others [7, 8, 9, 10]. However, there are still no preferred definitions. In this work, the definition adopted follows the formula proposed in [11]: 'A trust B to do X in a time T in a context C'; which adapted to our scenario would be: "*Annotator* trusts *subject in the video* in a time T and a *zero-acquaintance (first impression)* context".

Even with the main characteristics that define 'trust', the second issue in the field is how to measure it. According to [12], most of the studies still use methods from the '90s, such as lab games [13, 14].

These lab games allowed advance in science, first, through the development of several rules to estimate trust [15], which later evolved into the investigation of multimodal features and algorithms to understand which people's behaviors made them look more trustworthy against other subjects. An example is the study of G. Lucas et. al. [16], who introduced the terms objective trustworthiness (i.e., whether participants were being honest or not) and perceived trustworthiness (i.e. how participants were perceived by their game partner). They collected annotations on both trustworthiness perspectives and modeled them by automatically tracking the subjects' sentiments (positive or negative), expressiveness, and attention.

Despite the utility of lab games, trustworthiness should also be explored through other data, methods, and perspectives. Currently, some datasets with audio-visual material are appearing, such as those concentrated on deception recognition (whether the person is lying or not), by describing an object as in Box of Lies [17] or by counting lies in interactions with a partner, as in Deception Detection and Physiological Monitoring dataset [18]. Other corpora are oriented to understand perceived 'expertise'

10.21437/IberSPEECH.2022-42

of Conversational Agents [19], or to real-trials [20], where they extract fragments of real-trials videos and annotate data in terms of concordance with the story and the final sentence.

However, most of them still suffer from certain limitations such as the number of samples, the quality of the data (e.g. noisy recordings), or the context in which they are recorded that is subordinated to a specific task (e.g. describing an object). In this article, we study and analyze a methodology for reliably acquiring perceived trustworthiness annotations closer to a real scenario. We consider that this study could be relevant to learn more about the annotation process of this new type of material and understand the relationship between trustworthiness and other complex concepts such as voice attractiveness or eloquence. Additionally, examining new ways of annotating large amounts of data could benefit the training of deep learning models to automatically predict perceived trustworthiness, overcoming some current limitations in the field.

3. Materials and Methods

3.1. Annotation Procedure

First, YouTube was used to upload the chosen videos. The questions and responses for the surveys were then generated using an implemented Google App Script and an Excel document as a template. The forms were generated automatically when the script was run, then they were submitted to the annotators for online completion.

In total, we created two surveys, selecting videos of different actors representing in each video a single emotion. Each survey contained ten videos with ten queries per video, to filling by the annotators, related to certain characteristics of the subjects in the video, specifically: their familiarity, physical attractiveness, voice attractiveness, expressiveness, comfortableness in front of the camera, kindness, eloquence, the emotion displayed, the authenticity of the emotion, and the transmitted trustworthiness. Except for the familiarity, the authenticity of the shown emotion, and the trustworthiness questions, all of them were scored on a 7-point scale from -3 to 3. The other three only allowed Yes/No responses.

In addition to the questions related to the videos, we added a prior section to extract annotators' demographic data (age, level of education, and gender).

In this prior study, we sought to analyze the collected responses to identify the relationships across questions, the suitability of the approach to annotate emotional videos without inflicting bias, and the appropriateness of the answers' ranges.

3.2. Stimuli

The main corpus used for this project is RAVDESS [21], an open-source dataset for the research community, usually used in previous studies for emotion recognition [22, 23, 24]. This dataset includes 7356 recordings with acted expressions.

Only selecting the full Audio-Visual material and the speech channel, the dataset contains 1440 videos with a maximum and minimum duration of 5.31 and 2.99 s, respectively. The corpus has 24 actors distributed in a gender-balanced way, who speak lexically-matched statements in a neutral North American accent. This setup is suitable for studying the paralinguistics associated with trustworthiness, isolating the lexical. Also, each file contains a single actor representing one emotion, showing only their face and shoulders with a white background, without any element that could distract the observer's attention. The actors are between 20-40 years old, presenting different physical aspects and hairstyles.

We randomly selected 20 audio-visual clips of the RAVDESS dataset, guaranteeing that they were from actors 1 to 20, with distinct emotions displayed.

3.3. Participants

This research was conducted with students of the Universiti Sains Malaysia (Malaysia), consisting on 39 annotators (5 men and 34 women); 22 of them rated the survey 1 and 17, the survey 2. The annotators had a mean age of 22.85 (SD 2.12, 21-30). Of the annotators, 23 (58.97%) were studying the degree, 12 (30.77%) had completed the degree, and 4 (10.26%) had postgraduate studies (MS, MBA, JD, Ph.D., MD ...).

4. Results and Discussion

In this analysis, version 26.0 of the IBM SPSS Statistics for windows was employed to extract statistics from the responses retrieved from the questionnaires.

As the answers to the trustworthiness question were categorical (Yes/No), we applied nonparametric statistical metrics to study the scores assigned to this question.

First, we made an annotation analysis to understand the relationship between trustworthiness annotations and emotions.

After this, Spearman's correlations were extracted to examine the association between the trustworthiness question and the rest of the range-based questions of the form (including physical attractiveness, voice attractiveness, expressiveness, comfortableness in front of the camera, kindness, eloquence, and authenticity of the emotion displayed in the video). Later, we also calculated Spearman's correlation between the trustworthiness scores and the demographics of the annotators.

Finally, a feature selection procedure was applied to aural features, reporting the top 20 attributes more correlated with trustworthiness annotations.

4.1. Annotation Analysis

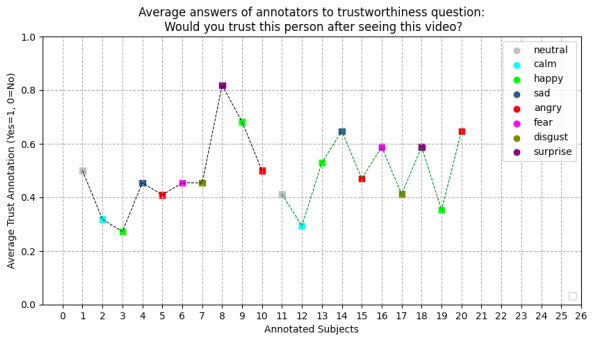

Figure 1: *Average trustworthiness scores per form. Color of the squares represents the ground-truth emotion displayed in the RAVDESS video.*

To analyze the relationship between the subjects in the video and their associated emotion, we assigned to each video the average trustworthiness score obtained across all the annotators, ranging from 0 (non-trustworthy person) to 1 (trustworthy person). The result of this conversion appears in Figure 1, which reveals that positive emotions like 'Happiness' do not always receive the highest scores despite previous studies' results that suggest that positive emotions correlate with trustworthi-

ness [25]. This non-reliability effect on 'Happiness' may be explained by the perception of the emotion as fake. In those cases where the subject's emotions are seen as acted, the trust in the person could be reduced (see video three). More interesting is video 8, in which the actor receives the top punctuation despite showing a non-positive valence emotion. This outcome encourages us to think that certain visual and vocal clues help to understand trust despite the expressed emotions.

4.2. Spearman's Correlations

4.2.1. Trustworthiness Perceived Characteristics Correlation

Figure 2 shows Spearman's correlation (ρ) matrix of trustworthiness question with other attributes rated in our forms about the person that appeared in the video, such as attractiveness or expressiveness.

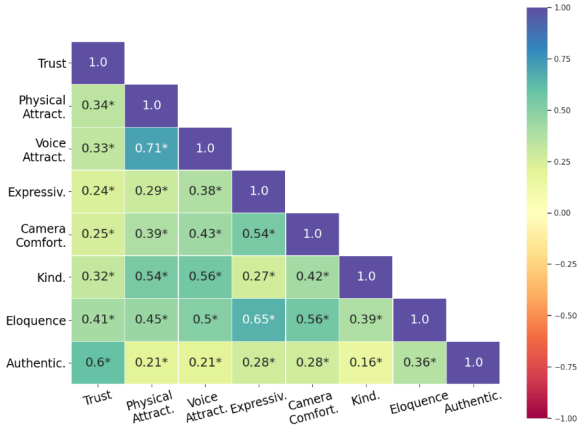

Figure 2: *Correlation between questions. (*) Indicates statistical significant ($p < 0.01$).*

Authenticity obtained the highest ρ = 0.6 ($p < 0.01$). This statistic may indicate that subjects showing authentic emotions received higher trustworthiness scores than others whose emotions seem simulated. To understand if the authenticity of the emotion could be related to errors in the perceived emotion, we also calculated the correlation between the errors at guessing emotions and the authenticity score, obtaining a significant ρ of 0.2. This ρ suggests that when some annotators fail to identify the emotion, probably they will consider it non-authentic. However, this low correlation also may indicate that other clues help to understand the authenticity score, motivating the idea that people are evaluating how genuine or credible the emotion is and not only how predictable it is.

Continuing with the analysis, we can see that eloquence, kindness, and attractiveness (physical and vocal) are also positively correlated with trustworthiness, getting ρ values above 0.3. Expressiveness and comfortableness in front of the camera have minor relationships with trustworthiness, but still, they are significant and positively correlated.

4.2.2. Audio Perceived Characteristics Correlation

On the voice-related characteristics, notice that question of eloquence has a strong correlation with expressiveness ($\rho = 0.65$) and voice attractiveness ($\rho = 0.5$), two characteristics usually associated with eloquence.

Regarding the voice attractiveness correlation, it seems that voice and physical attractiveness also show a high positive relationship. This effect could have several implications: first, it could be occasioned by the order in which questions appear, which makes people tend to annotate both questions in the same manner; or, second, that one of them influences the judgments on the other, which would mean that people with pleasant voices seem more physically attractive, or people with beautiful faces are perceived as with genuine vocal tones too. More studies should be made to corroborate these hypotheses. Regarding the rest of the dimensions, there is also a high correlation between kindness ($\rho = 0.56$) and comfortableness in front of the camera ($\rho = 0.43$). The high correlation of kindness with voice attractiveness suggests that kindness could be modeled using vocal clues, which could be an interesting future study. The camera comfortableness correlation score indicates that the more relaxed people look, the more attractive their voices are.

4.2.3. Annotators Demographics Correlations

Although we performed correlations between different demographics of the participants (Age, Educational Level, and Gender) with their trustworthiness scores, only the Educational Level reported a significant ρ value of -0.19. This negative correlation might imply that people with university studies tend to trust unknown people less than others without studies (at least in absence of prior information about the person).

4.2.4. Survey Correlations

Another concern when creating a dataset is the order in which information is shown to the annotator.

To study the effect that the order of the videos may have in the annotation process, we calculated the correlation between the video's position and trustworthiness score, getting a significant ρ=0.14 ($p < 0.01$). Although the correlation is low, this result may indicate that the video's position could affect the trustworthiness of annotations. For this reason, in future versions of the surveys, we will randomize the order of the subjects and emotions.

4.3. Attribute selection

To conclude the experimentation phase, we performed an attribute selection experiment on aural features, using the estimated trustworthiness level as labels. This trustworthiness ground-truth of each video was calculated as the average of all the users' annotations on that video.

Regarding the aural features, we used the eGeMAPS set [26]. This set contains 88 statistical functional features (average, standard deviation, percentiles, etc.) obtained from aural Low-Level Descriptors (LLD) that belonged to 5 different families: Frequency-related, Energy/Amplitude-related, Spectral-related, Temporal-Related, and Quality. The reason for using this set is that it has been employed in several affective-computing tasks as a baseline [27, 28]. Therefore, as trustworthiness could also be a psychological state, these features could be suitable to establish a baseline and comprehend which families of attributes could help to explain perceived trustworthiness in our context of emotional videos.

Regarding the validation strategy, we used Leave-One-Out to utilize each user at least once as a test set, ensuring that the selected attributes were the most reliable for predicting trustworthiness in a general case.

For performing feature selection, we employed the implemented CorrelationAttributeEval method of Weka 3.8.3 [29]. This algorithm evaluates the correlation between each attribute

# Features eGeMAPS per family	Name of Attributes per Family	# Top 20 Selected
	Frequency related	**11/20**
	Pitch	5
	Jitter	2
24/88	F1 frequency	1
	F2 frequency	1
	F3 frequency	1
	F3 bandwidth	1
14/88	**Energy/Amplitude related**	**1/20**
	Harmonics-to-Noise Ratio (HNR)	1
	Spectral (balance/shape/dynamics)	**6/20**
	mfcc1	1
	mfcc2	1
43/88	Voiced segments	
	mfcc4	1
	Alpha Ratio - Voiced Segments	1
	Spectral Slope 0–500 Hz	1
	Voiced Segments	
	Hammarberg Index	1
6/88	**Temporal**	**1/20**
	Voiced Segments per Second	1
1/88	**Quality**	**0/20**
-	**Other**	**1/20**
	Emotion	1

Table 1: *Top 20 features from the eGeMAPS set selected using the CorrelationAttributeEval feature selector of Weka.*

and the class. Hence, the most correlated features appear in the leading positions of the ranking. Table 1 shows the top 20 over 88 attributes of eGeMAPS that achieved the highest Pearson's correlation. As we can see, 11 features belonged to the Frequency-related family, followed by the Spectral family with six characteristics, and the Temporal and Energy/Amplitude groups with one attribute each. These results suggest that certain behaviors like the speaking velocity (voiced segments per second) or changes in the pitch or F1, F2, and F3 frequencies could carry relevant information about how trustworthy a person is perceived.

Additionally, when emotion is included in the set of aural features, it is also selected as a relevant attribute for understanding perceived trustworthiness. This fact agrees with previous studies [16] and confirms the necessity of not mixing emotions when the target is to reduce its effects on trustworthiness annotations. Using forms with single emotions or calculating the average of mixed-emotional surveys could help to remove the emotional bias to study other relevant trustworthiness indicators.

5. Limitations

As this is a preliminary study, one of the main limitations is the number of samples, only 20 subjects were evaluated in a chosen emotional state. For future versions, we will include the same users in different emotional states and random orders to reduce the bias that emotions and order introduce in the annotation process. As a result of this change, we could calculate a more reliable averaged trustworthiness score for the subjects across emotions.

Regarding the survey questions and answers, we will change the dichotomous answers to a range-based option to apply parameter-based statistics and compare all the questions in the same ranges, avoiding possible deviations in the scores due to distinct answer types.

We also noticed that the agreement in most videos was low, which could be caused by the different dimensions that trust, as a concept, has. Maybe some raters were not thinking the same when we asked about the trustworthiness raised by the subject in the video. For example, some raters may think about trusting the person to take care of their plants and others to take care for their children. To reduce divergences, we will add some context to the question to ensure that all the annotators have a more accurate idea of what trustworthiness aspect they need to annotate.

6. Conclusions

Despite the earlier mentioned limitations, this study has also come up with several findings in trustworthiness annotation.

First, trustworthiness scores show that different subjects lead to higher agreements than others, which suggests that it is possible to measure differences between individuals in short videos at zero-acquaintance, despite the subjective nature of the task. It implies that people are sensitive to different characteristics of the subjects, but also that our methodology can measure and quantify some of these differences.

Next, we discovered the relationships between trustworthiness and other relevant features for Malaysian annotators from the correlation analysis. One of these features is the authenticity of the acted emotions and their importance even in negative emotions. This fact is supported by previous studies on types of smiles, in which genuine smiles received higher scores than others [16]. But also, other features such as expressiveness correlate with trustworthiness judgments, which are also in line with previous results in the field [30].

Other features, such as eloquence or kindness, also reveal a high correlation with trustworthiness which opens an interesting future line of study. The same happens with voice and physical attractiveness (with a ρ=0.71), which we will study further by isolating audio and visual content and changing the order of the form questions.

Regarding attribute selection, the frequency-related features family was the most chosen, which suggests that pitch, jitter, and certain frequency variations affect trustworthiness perception in people's voices. In the future, these features could be fed to machine learning algorithms to learn to predict trustworthiness automatically.

Finally, as future lines, we will extend the study to more subjects to generate a larger dataset to perform trustworthiness recognition; and analyze cross-cultural scenarios to understand if people from other countries have the same perception of trustworthiness.

7. Acknowledgements

The work leading to these results was supported by University Sains Malaysia from the grant no. 304/PKOMP/6315137 and the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C21 and PID2020-118112RB-C22, funded by MCIN/AEI/10.13039/501100011033), and AMIC-PoC (PDC2021-120846-C42, funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR").

We also thank the 'Ayudas Programa Propio UPM 2022', and the Ricardo Kleinlein's research FPI grant PRE2018-083225 given by the Spanish Ministry of Education.

8. References

[1] J. Brennan, "Recognition trust," *Philosophical Studies*, vol. 178, no. 11, pp. 3799–3818, Mar. 2021. doi: 10.1007/s11098-021-01626-z

[2] A. Kumar, V. Capraro, and M. Perc, "The evolution of trust and trustworthiness," *Journal of The Royal Society Interface*, vol. 17, no. 169, p. 20200491, Aug. 2020. doi: 10.1098/rsif.2020.0491

[3] F. Kocsor, L. Kozma, A. L. Neria, D. N. Jones, and T. Bereczkei, "Arbitrary signals of trustworthiness – social judgments may rely on facial expressions even with experimentally manipulated valence," *Heliyon*, vol. 5, no. 5, p. e01736, 2019. doi: 10.1016/j.heliyon.2019.e01736

[4] M. Weymar, C. Ventura-Bort, J. Wendt, and A. Lischke, "Behavioral and neural evidence of enhanced long-term memory for untrustworthy faces," *Scientific Reports*, vol. 9, no. 1, dec 2019. doi: 10.1038/s41598-019-55705-7

[5] C. Wienrich, C. Reitelbach, and A. Carolus, "The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of voice assistants, providers, data receivers, and automatic speech recognition," *Frontiers in Computer Science*, vol. 3, 2021. doi: 10.3389/fcomp.2021.685250

[6] V. Franzoni, A. Milani, and J. Vallverdú, "Emotional affordances in human-machine interactive planning and negotiation," in *Proceedings of the International Conference on Web Intelligence*, ser. WI '17. New York, NY, USA: Association for Computing Machinery, 2017. doi: 10.1145/3106426.3109421 p. 924–930.

[7] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of Management Review*, vol. 23, no. 3, pp. 393–404, Jul. 1998. doi: 10.5465/amr.1998.926617

[8] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Trans. Comput.-Hum. Interact.*, vol. 12, no. 2, p. 293–327, jun 2005. doi: 10.1145/1067860.1067867

[9] M. Rheu, J. Y. Shin, W. Peng, and J. Huh-Yoo, "Systematic review: Trust-building factors and implications for conversational agent design," *International Journal of Human–Computer Interaction*, vol. 37, no. 1, pp. 81–96, 2021. doi: 10.1080/10447318.2020.1807710

[10] J. Li, B. Mao, Z. Liang, Z. Zhang, Q. Lin, and X. Yao, "Trust and trustworthiness: What they are and how to achieve them," in *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2021. doi: 10.1109/PerComWorkshops51409.2021.9430929 pp. 711–717.

[11] E. M. Uslaner, Ed., *The Oxford Handbook of Social and Political Trust*. Oxford University Press, Jan. 2017.

[12] ——, *The Oxford Handbook of Social and Political Trust*. Oxford University Press, Jan. 2017.

[13] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity, and social history," *Games and Economic Behavior*, vol. 10, no. 1, pp. 122–142, 1995. doi: https://doi.org/10.1006/game.1995.1027

[14] P. C. Bauer and M. Freitag, *Measuring Trust*, E. M. Uslaner, Ed. Oxford University Press, Jan. 2017.

[15] Q.-V. Dang and C.-L. Ignat, "Computational trust model for repeated trust games," in *2016 IEEE Trustcom/BigDataSE/ISPA*, 2016. doi: 10.1109/TrustCom.2016.0043 pp. 34–41.

[16] G. Lucas, G. Stratou, S. Lieblich, and J. Gratch, "Trust me: Multimodal signals of trustworthiness," ser. ICMI '16. New York, NY, USA: Association for Computing Machinery, 2016. doi: 10.1145/2993148.2993178 p. 5–12.

[17] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "Box of lies: Multimodal deception detection in dialogues," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019. doi: 10.18653/v1/N19-1175 pp. 1768–1777.

[18] J. Speth, N. Vance, A. Czajka, K. W. Bowyer, D. Wright, and P. Flynn, "Deception detection and remote physiological monitoring: A dataset and baseline experimental results," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 2021. doi: 10.1109/IJCB52358.2021.9484409 pp. 1–8.

[19] L. Gauder, P. Riera, L. Pepino, S. Brussino, J. Vidal, L. Ferrer, and A. Gravano, "Trust-uba: A corpus for the study of the manifestation of trust in speech," 2020.

[20] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: Association for Computing Machinery, 2015. doi: 10.1145/2818346.2820758 p. 59–66.

[21] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," April 2018. doi: 10.5281/zenodo.1188976 Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak.

[22] H. A. Abdulmohsin, H. B. Abdul wahab, and A. M. J. Abdul hossen, "A new proposed statistical feature extraction method in speech emotion recognition," *Computers & Electrical Engineering*, vol. 93, p. 107172, 2021. doi: 10.1016/j.compeleceng.2021.107172

[23] M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla, and C. Benavides, "Sentiment analysis in non-fixed length audios using a fully convolutional neural network," *Biomedical Signal Processing and Control*, vol. 69, p. 102946, 2021. doi: 10.1016/j.bspc.2021.102946

[24] P. Singh, R. Srivastava, K. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowledge-Based Systems*, vol. 229, p. 107316, 2021. doi: 10.1016/j.knosys.2021.107316

[25] J. R. Dunn and M. E. Schweitzer, "Feeling and believing: The influence of emotion on trust." *Journal of Personality and Social Psychology*, vol. 88, no. 5, pp. 736–748, 2005. doi: 10.1037/0022-3514.88.5.736

[26] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016. doi: 10.1109/TAFFC.2015.2457417 Open access.

[27] Y. Matveev, A. Matveev, O. Frolova, E. Lyakso, and N. Ruban, "Automatic speech emotion recognition of younger school age children," *Mathematics*, vol. 10, no. 14, 2022. doi: 10.3390/math10142373

[28] L. Ilias and D. Askounis, "Multimodal deep learning models for detecting dementia from speech and transcripts," *Frontiers in Aging Neuroscience*, vol. 14, 2022. doi: 10.3389/fnagi.2022.830943

[29] E. Frank, M. A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, and I. H. Witten, *Weka: A machine learning workbench for data mining*. Berlin: Springer, 2005, pp. 1305–1314.

[30] R. T. Boone and R. Buck, *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 163–182, 2003. doi: 10.1023/a:1025341931128