2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

COVID-19: Affect recognition through voice analysis during the winter lockdown in Scotland

Sofia de la Fuente Garcia, Fasih Haider and Saturnino Luz

Abstract— **The COVID-19 pandemic has led to unprecedented restrictions in people's lifestyle which have affected their psychological wellbeing. In this context, this paper investigates the use of social signal processing techniques for remote assessment of emotions. It presents a machine learning method for affect recognition applied to recordings taken during the COVID-19 winter lockdown in Scotland (UK). This method is exclusively based on acoustic features extracted from voice recordings collected through home and mobile devices (i.e. phones, tablets), thus providing insight into the feasibility of monitoring people's psychological wellbeing remotely, automatically and at scale. The proposed model is able to predict affect with a concordance correlation coefficient of 0.4230 (using Random Forest) and 0.3354 (using Decision Trees) for arousal and valence respectively.**

Clinical relevance— **In 2018/2019, 12% and 14% of Scottish adults reported depression and anxiety symptoms. Remote emotion recognition through home devices would support the detection of these difficulties, which are often underdiagnosed and, if untreated, may lead to temporal or chronic disability.**

I. Introduction

The Coronavirus Disease 2019, commonly known as COVID-19, is a viral respiratory syndrome caused by a highly infectious, novel strain of the coronavirus family (SARS-CoV-2) [1]. It emerged at the end of 2019 and rapidly became a pandemic, having an unprecedented impact across the world, with nearly 200 million cases and over 4 million deaths in July 2021 [2].

In Scotland, the COVID-19 lockdown was announced in March, 2020. The restrictions entailed a ban on all non-essential travel and activities and the advice to work from home and stay at home, in order to slow the spread of the disease and reduce the burden on the National Health Service (NHS). Consequently, everyone's social and professional lives were seriously disrupted, masks became mandatory, "self-isolation" and "social distancing" made their way into our daily vocabulary and people had to adapt to being confined to their homes. After the summer, cases peaked again and the Scottish government announced a second lockdown just as days were getting darker and colder.

Also in March, the World Health Organisation voiced concerns over the impact of the pandemic on global mental health and published guidelines for self-care, stress, fear management, and dealing with self-isolation [3]. Initial research evidence suggests an increased incidence of anxiety and depression in this period due to persisting stay-at-home orders, financial insecurity, loneliness and overall uncertainty [4].

Early in the pandemic, a Chinese study found a significant drop in life satisfaction and an increased anxiety in 18,000 social media users [5]. Furthermore, 54% of the population in China reported that COVID-19 had a moderate to severe impact on their well-being [6], which was more pronounced in healthcare professionals [7]. Later studies reported similar trends in Europe, America and Oceania. A German study found that 50% of the population suffered from anxiety and spent several hours per day thinking about the pandemic [8]. In Spain, an online study with 3,480 people found increased rates of depression (18.7% participants), anxiety (21.6%) and post-traumatic stress disorder (PTSD; 15.8%), for which loneliness was the strongest risk factor [9]. A Brazilian study also found that 20% of their participants reported severe distress [10]. In the U.S., a survey found increased levels of anxiety and depression as well as increasing financial and health concerns, with especially severe implications for older adults [11]. In New Zealand, a comparable survey found substantial percentages of respondents experienced various degrees of psychological distress. Differently to the U.S. study, this survey found young people to be particularly affected, alongside those who had lost their jobs or had a past history of mental difficulties [12]. These studies show evidence that isolation is a risk-factor for mental health difficulties, especially depression and anxiety.

The study reported here took place in Scotland, where depression and anxiety were already a major health concern before the COVID-19 outbreak. In 2016, the Scottish Burden of Disease Study [13] estimated nearly half a million adults over 16 to be suffering from disability due to depression or anxiety. More recently, in the Scottish Health Survey 2018/2019, 12% and 14% of Scottish adults reported at least two depression and anxiety symptoms, respectively [14].

These difficulties are linked with life satisfaction as they reduce one's ability to feel joy, take pleasure in our activities and have the energy to fulfil our intentions and live a meaningful life [15]. Isolated people are also less likely to receive help for the difficulties they experience [16]. In this context, our work focuses on emotion recognition through voice analysis, in order to assess the feasibility of mental e-health systems. We present an automatic approach that consists of extracting acoustic features from spontaneous speech data collected during the winter lockdown and training a machine learning model to predict participants' emotions and energy levels.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

All authors are with the Usher Institute, in the Edinburgh Medical School. The University of Edinburgh, UK. sofia.delafuente@ed.ac.uk

II. Related Work

Most work on machine emotion recognition has been done through image and facial processing [17]. However, emotion recognition through speech analysis is progressively gaining momentum [18], especially so in the context of health research [19]. Speech and language carry information about the speaker, including age and gender [20], as well as physiological, behavioural and emotional information [21]. Speech is ubiquitous and may be collected automatically, unobstrusively and with relatively little infrastructure. This has led to increasing research on technologies for personal health monitoring and diagnostic support tools based on automated processing of speech, in which the field of emotion recognition has been especially active over the past decade. These technologies are supported by machine learning and artificial intelligence, which enable a broad range of analysis and recognition tasks [22].

The premise is that, under different emotions, we produce the same utterance with detectable acoustic differences [23]. Since mental health difficulties often progress with emotional changes, there is an opportunity for signal processing to support the detection of depression, anxiety, apathy, and suicidality [24]. For instance, a recent study on adolescents was able to recognise suicidal ideation thorough an analysis of their voice, using only acoustic features [25]. Another recent study found similar results in their attempt to identify suicidal ideation in war veterans [26].

Emotional processing is challenging, partially because of emotional complexity and a degree of overlap across the emotions that are conventionally designated as primary (sadness, happiness, fear, anger, disgust and surprise) [27]. Categorical recognition of these primary emotions through speech has been a major research focus [28], [29], [30], [31], but a dimensional approach is increasingly popular. This considers primary emotions to be interdependent and operationalises them in two dimensions, namely Valence (V) and Arousal (A) [32], [33], [34], [35]. Briefly, Valence refers to how positive or negative an emotional state is, from unpleasant to pleasant; while Arousal refers to how excited and activated such emotional state feels.

In this work we conceptualise emotionality and affect in these two dimensions (A and V) in order to capture the subtleties that may be overlooked by single-emotion labels. Scores for both of these dimensions were obtained by self-rating through an affective slider [36]. We analysed the voices of 109 individuals and built models to predict their affect scores based on their speech, collected during the winter lockdown. Additionally, we analysed these scores in relation to the Hospital Anxiety and Depression Scale (HADS) [37] in order to establish their association with this symptomatology. To the best of our knowledge, this is the first study utilising speech analysis for recognition of emotional dimensions that are supported with a clinically validated tool (HADS).

III. Method

A. Dataset

The PsyVoiD[1] project investigates the relationship between speech and well-being in the context of the COVID-19. The present subset contains 109 participants who completed the study in Scotland during the winter lockdown.

Participants were between 26 and 86 years old ($\overline{X} = 59$). Sixty nine of them are female (63%) and 34 had a past history of depression (31%). All of them completed the HADS questionnaire, which yields anxiety ($M = 6$) and depression scores ($M = 4$). The pleasure and arousal scales range from $-100 to 100$ ($\overline{X} = 34.54$ and 30.17, respectively). Table I presents these descriptive statistics.

TABLE I: Descriptive statistics on the 109 subjects.

Variable	Value
Average age (std)	59.23 (12.44)
Gender (F/M)	69/41
Past history of depression (Y/N)	34/71
HADS Anxiety score (M)	6
HADS Depression score (M)	4
Arousal score ($\overline{X}$, std)	30.17 (52.44)
Pleasure score ($\overline{X}$, std)	34.54 (54.28)

We report the median (M) for HADS due to its skewed distribution (i.e. as this is a non-clinical population, the majority of participants have low scores).

B. Correlation Analysis

The prompt used to collect the affective scores is shown in Figure 1. In order to support the clinical validity of the valence and arousal measures, we performed a Pearson's product-moment correlation between them and the HADS scores. Both were negatively and significantly correlated with HADS ($p < 0.01$), with coefficients of -0.62 and -0.71, respectively. Figure 2 represents this association. Hence, participants with higher scores on the HADS questionnaire reported lower levels in the affective slider.

Fig. 1: Affective slider: Arousal (top):'How "energised" do you feel right now?'; and Pleasure (bottom): 'How pleased do your feel at the moment?'.

C. Speech processing

The audio files were pre-processed in order to ensure consistency. We implemented spectral substraction for stationary noise removal, audio volume normalisation to control for variable recording conditions, and voice activity detection (VAD) based on signal energy thresholding. The resulting enhanced and segmented recordings were used for acoustic feature extraction and machine learning prediction, and may be made available upon request. There are 3,242 segments from 109 recordings, with an average duration of 72.29 seconds ($sd = 27.03$).

[1]Ethical approval: June 15, 2020 (REC reference: 20/EM/0146)

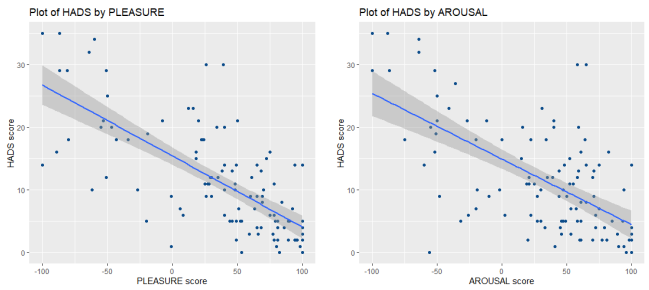

Fig. 2: Association between HADS and Affect scores (shaded are indicates confidence interval).

D. Acoustic feature extraction: eGeMAPS and ADR

The resulting 3,242 acoustic segments were used to extract a comprehensive paralinguistic feature set: *eGeMAPS* [38]. This feature set contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index and slope V0 features, as well as their most common statistical functionals, totalling 88 features per 100ms frame. Features were chosen given their theoretical significance and potential to detect physiological changes in voice production. The *eGeMAPS* set has also proven useful for detection of medical conditions in previous studies [39].

Using the segment level acoustic information extracted, we applied the active data representation method (ADR) to generate a data representation for each audio recording. ADR employs self-organising maps to cluster the original acoustic features and then computes second-order features over these to extract new features. It has been tested previously for large scale time-series data (see [39] for details).

E. Regression Analysis

We used five types of regression models, namely, linear regression (LR), Decision Trees (DT; where leaf size is optimised through grid search between 1 and 20 and CART algorithm), support vector regression (SVR; with a linear kernel, box constraint k optimised though grid search over $k \in \{10^5, 20^5, ...10^6\}$, and sequential minimal optimisation solver), Random Forest regression ensembles (RF; where leaf size is optimised through grid search between 1 and 20 with 10 trees in the forest), and Gaussian process regression (GP; with a squared exponential kernel). The regression methods were implemented in MATLAB.

IV. Results and Discussion

We evaluated a model based on a comprehensive paralinguistic feature set (*eGeMAPS*), a non-linear method for feature extraction (ADR) and five machine learning algorithms for regression. We used the concordance correlation coefficient (CCC) to measure the agreement between the target and predicted scores. CCC is commonly used in emotion recognition and is effectively a non-linear combination of Pearson's correlation coefficient and the mean square error [40]. The results are summarised in Table II.

Our model was able to predict affect with the best CCC of 0.4230 (RF) and 0.3354 (DT) for A and V, respectively. These CCC scores are lower than results reported elsewhere, such as the SEMAINE dataset (0.680 and 0.506 CCC for A and V, respectively) and the RECOLA dataset (0.692 and 0.423) [41]. However, we note that those datasets consist of recordings taken in carefully controlled conditions and annotated by groups of 6+ annotators, which ensured a level of uniformity unavailable in real-life, self-rated data.

While previous studies typically attempted affect recognition at short ($\approx 20s$) speech segment and/or sentence level [29], [30], [31], this study proposes a system to recognise affect by generating a representation for an entire recording, using segment level information and our ADR method. To our best knowledge, this is the first approach to dimensional affect recognition system using acoustic information for large audio segments ($\overline{X} = 72.29$ seconds, $sd = 27.03$).

Another relevant limitation of previous studies is the size of the available dataset. There are 10 participants in the EmoDB [29], 4 in the SAVEE dataset [30], 6 in the EMOVO dataset[31], 10 in the vlogger dataset [42]), 23 in the SEMAINE [34] and 46 in the RECOLA [35]. Our study, on the other hand, contains speech of 109 participants and 3,242 segments.

Finally, emotion recognition is often limited by the inherent subjectivity in having emotions labelled by humans who perceive affect from audio, visual and linguistic information [43]. In this study, this intermediate step was removed as the affect is self-reported through the affective slider. In addition, the affective scores were validated by their statistically significant association with the HADS questionnaire ($p < 0.01$, $\rho_{Arousal} = -0.62$ and $\rho_{Valence} = -0.71$), a long standing tool to screen for depression and anxiety.

TABLE II: Leave-one-subject out cross-validation results (CCC)

Task	LR	DT	SVM	RF	GP
Arousal	0.1974	0.3841	0.3389	**0.4230**	0.0215
Valence	0.1436	**0.3354**	0.3131	0.3256	0.0978

V. Conclusions

The COVID-19 pandemic has caused great psychological distress among the population. Remote emotion recognition through home devices could support the detection of these difficulties, which are often underdiagnosed and, if untreated, may lead to temporal or chronic disability. Timely identification of these difficulties might enhance the effect of any therapeutic intervention.

This study proposes an affect recognition method with a view to contributing to automatic and remote monitoring of emotional health. Our results evidence the potential to detect emotional and mood difficulties through spontaneous spoken language. In addition, by employing solely on acoustic features, our method is language-independent, and potentially privacy preserving.

Future work will involve analysing linguistic information and its fusion with acoustic information. Other psychological variables available in the dataset will be explored in order to comprehensively assess the potential of our system for mental e-Health.

REFERENCES

[1] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu *et al.*, "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," *The lancet*, vol. 395, no. 10224, pp. 565–574, 2020.

[2] World Health Organisation, "WHO: Coronavirus disease (COVID-19) dashboard." [Online]. Available: https://covid19.who.int/

[3] ——, "Mental health and psychosocial considerations during the covid-19 outbreak, 18 march 2020," 2020.

[4] N. Vindegaard and M. E. Benros, "Covid-19 pandemic and mental health consequences: Systematic review of the current evidence," *Brain, behavior, and immunity*, vol. 89, pp. 531–542, 2020.

[5] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, "The impact of covid-19 epidemic declaration on psychological consequences: a study on active weibo users," *International journal of environmental research and public health*, vol. 17, no. 6, p. 2032, 2020.

[6] C. Wang, R. Pan, X. Wan, Y. Tan, L. Xu, C. S. Ho, and R. C. Ho, "Immediate psychological responses and associated factors during the initial stage of the covid-19 epidemic among the general population in china," *International journal of environmental research and public health*, vol. 17, no. 5, p. 1729, 2020.

[7] J. Lai, S. Ma, Y. Wang, Z. Cai, J. Hu, N. Wei, J. Wu, H. Du, T. Chen, R. Li *et al.*, "Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019," *JAMA network open*, vol. 3, no. 3, pp. e203 976–e203 976, 2020.

[8] M. B. Petzold, A. Bendau, J. Plag, L. Pyrkosch, L. Mascarell Maricic, F. Betzler, J. Rogoll, J. Große, and A. Ströhle, "Risk, resilience, psychological distress, and anxiety at the beginning of the COVID-19 pandemic in Germany," *Brain and behavior*, vol. 10, no. 9, 2020.

[9] C. González-Sanguino, B. Ausín, M. Á. Castellanos, J. Saiz, A. López-Gómez, C. Ugidos, and M. Muñoz, "Mental health consequences during the initial stage of the 2020 coronavirus pandemic (covid-19) in spain," *Brain, behavior, and immunity*, vol. 87, pp. 172–176, 2020.

[10] S. Zhang, Y. Wang, A. Jahanshahi, and V. Schmitt, "First study on mental distress in brazil during the COVID-19 crisis," *medRxiv*, 2020.

[11] W. Bruine de Bruin, "Age differences in covid-19 risk perceptions and mental health: Evidence from a national us survey conducted in march 2020," *The Journals of Gerontology: Series B*, vol. 76, no. 2, pp. e24–e29, 2021.

[12] S. Every-Palmer, M. Jenkins, P. Gendall, J. Hoek, B. Beaglehole, C. Bell, J. Williman, C. Rapsey, and J. Stanley, "Psychological distress, anxiety, family violence, suicidality, and wellbeing in new zealand during the covid-19 lockdown: A cross-sectional study," *PLoS one*, vol. 15, no. 11, p. e0241658, 2020.

[13] Public Health informatoin for Scotland (ScotPHO), "The Scottish Burden of Disease Study, 2016: Depression technical overview," 2016.

[14] Scottish Government, "Scottish Health Survey," 2018.

[15] B. Headey, J. Kelley, and A. Wearing, "Dimensions of mental health: Life satisfaction, positive affect, anxiety and depression," *Social indicators research*, vol. 29, no. 1, pp. 63–82, 1993.

[16] J. T. Cacioppo and L. C. Hawkley, "Social isolation and health, with an emphasis on underlying mechanisms," *Perspectives in biology and medicine*, vol. 46, no. 3, pp. S39–S52, 2003.

[17] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.

[18] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, 2017, pp. 109–114.

[19] S. J. Brown, "Multi-user remote health monitoring system," Aug. 8 2000, uS Patent 6,101,478.

[20] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Convolutional-recurrent neural network for age and gender prediction from speech," in *2019 Signal Processing Symposium (SPSympo)*. IEEE, 2019, pp. 242–245.

[21] H. Kaya, A. A. Salah, A. Karpov, O. Frolova, A. Grigorev, and E. Lyakso, "Emotion, age, and gender classification in children's speech by humans and machines," *Computer Speech & Language*, vol. 46, pp. 268–283, 2017.

[22] N. Cummins, A. Baird, and B. W. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[23] J. D. Williamson, "Speech analyzer for analyzing frequency perturbations in a speech pattern to determine the emotional state of a person," Feb. 27 1979, uS Patent 4,142,067.

[24] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[25] C. Figueroa Saavedra, T. Otzen Hernández, C. Alarcón Godoy, A. Ríos Pérez, D. Frugone Salinas, and R. Lagos Hernández, "Association between suicidal ideation and acoustic parameters of university students' voice and speech: a pilot study," *Logopedics Phoniatrics Vocology*, pp. 1–8, 2020.

[26] V. Sourirajan, A. Belouali, M. A. Dutton, M. Reinhard, and J. Pathak, "A machine learning approach to detect suicidal ideation in us veterans based on acoustic and linguistic features of speech," *arXiv preprint arXiv:2009.09069*, 2020.

[27] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *Journal of nonverbal behavior*, pp. 1–28, 2019.

[28] E. Mower, M. J. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2010.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of the ninth European Conference on Speech Communication and Technology*, 2005, pp. 1516–1520.

[30] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP)*, Sept. 2009, pp. 53–58.

[31] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, ser. LREC 2014. European Language Resources Association (ELRA), 2014, pp. 3501–3504.

[32] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[33] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

[34] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2011.

[35] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *10th international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[36] A. Betella and P. F. Verschure, "The affective slider: A digital self-assessment scale for the measurement of human emotions," *PloS one*, vol. 11, no. 2, p. e0148037, 2016.

[37] A. S. Zigmond and R. P. Snaith, "The hospital anxiety and depression scale," *Acta psychiatrica scandinavica*, vol. 67, no. 6, 1983.

[38] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "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, 2016.

[39] F. Haider, S. de la Fuente, and S. Luz, "An assessment of paralinguistic acoustic features for detection of alzheimer's dementia in spontaneous speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 272–281, 2020.

[40] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE transactions on affective computing*, vol. 8, no. 3, pp. 314–327, 2016.

[41] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks." in *INTERSPEECH*, 2018, pp. 3092–3096.

[42] F. Haider and S. Luz, "Attitude recognition using multi-resolution cochleagram features," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 3737–3741.

[43] F. Haider, S. Pollak, P. Albert, and S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods," *Computer Speech & Language*, vol. 65, p. 101119, 2021.