A Fala como Biomarcador de Carga Emocional em Aplicações Clínicas

Speech as an Emotional Load Biomarker in Clinical Applications

L. F. Coelho

Resumo:

Introdução: Os profissionais de saúde enfrentam frequentemente cargas emocionais significativas no seu trabalho, incluindo o impacto de emoções negativas, como o *stress* e a ansiedade, que podem ter graves consequências no desempenho das suas funções de prestação de cuidados de saúde imediatos e também na sua própria saúde a longo prazo. Neste artigo, é proposto um algoritmo de estimativa do *stress* baseado na classificação de emoções de valência negativa em gravações de fala.

Métodos: É proposto um *pipeline* de aprendizagem automática de ponta a ponta. São considerados cenários de modelos de decisão distintos, o VGG-16 e o SqueezeNet, que partilham uma entrada comum de espetrograma de potência Q constante para representação acústica. Os sistemas são treinados e avaliados utilizando os conjuntos de dados de fala emocional RAVDESS e TESS.

Resultados: O sistema foi avaliado para a classificação de um conjunto de emoções (problema multiclasse) e também para a classificação de emoções negativas e neutras, distinguindo-as das positivas (problema binário). Os resultados obtidos são comparáveis aos dos sistemas anteriormente registados, com o modelo SqueezeNet a oferecer uma pegada significativamente mais pequena, permitindo aplicações versáteis. Uma exploração mais aprofundada do espaço de parâmetros do modelo não foi exaustiva e por isso é promissora para a melhoria do desempenho.

Conclusão: O sistema proposto pode constituir uma abordagem viável para a estimativa de um biomarcador não-invasivo de baixo custo para emoções negativas. Isto permite ativar alertas e desenvolver ações de mitigação para a presença de emoções negativas, sendo uma ferramenta de gestão adicional para os serviços de saúde que permite manter a qualidade e maximizar a sua disponibilidade.

Palavras-chave: Aprendizagem Automática; Biomarcadores; Emoções; Fala.

Abstract:

Introduction: Healthcare professionals often contend with significant emotional burdens in their work, including the impact of negative emotions, such as stress and anxiety, which can have profound consequences on immediate and long-term healthcare delivery. In this paper a stress estimation algorithm is proposed based on the classification of negative valence emotions in speech recordings.

Methods: An end-to-end machine learning pipeline is proposed. Two distinct decision models are considered, VGG-16 and SqueezeNet, while sharing a common constant Q power spectrogram input for acoustic representation. The system is trained and evaluated using the RAVDESS and TESS emotional speech datasets.

Results: The system was evaluated for individual emotion classification (multiclass problem) and also for negative and neutral or positive emotion classification (binary problem). The results achieved are comparable to previously reported systems, with the SqueezeNet model offering a significantly smaller footprint, enabling versatile applications. Further exploration of the model's parameter space holds promise for enhanced performance.

Conclusion: The proposed system can constitute a feasible approach for the estimation of a low-cost non-invasive biomarker for negative emotions. This allows to raise alerts and develop mitigating actions to the burden of negative emotions, being an additional management tool for healthcare services that allows to maintain quality and maximize availability.

Keywords: Biomarkers; Emotions; Machine Learning; Speech.

Introduction

In healthcare facilities, medical professionals frequently encounter a variety of stressors due to the demanding nature of their work. These stressors encompass patient care pressures, including the responsibility for accurate diagnoses and critical treatment decisions, as well as the emotional toll of witnessing patients' suffering. Medical staff also contend with substantial workloads, time constraints, and the challenges of effective communication, along with administrative burdens and ethical dilemmas. Additionally, job security concerns, interpersonal conflicts, and the fear of medical errors can contribute to anxiety and other types of negative emotions. Over time, providing empathetic care may also lead to compassion fatigue. Healthcare workers must navigate all these emotions while still maintaining patient safety and advocating for their needs.

INESC TEC – Instituto de Sistemas e Computadores and ISEP / P.Porto – Instituto Superior de Engenharia do Porto / Instituto Politécnico do Porto, Porto, Portugal

https://doi.org/10.24950/rspmi.2587

The emotionally strong context faced by medical professionals in healthcare facilities carries several associated risks. Negative emotions, especially when prolonged in time, can lead to burnout, affecting job satisfaction and retention, while impairing cognitive function and empathy, potentially compromising patient safety and satisfaction. Mental health issues and physical health problems may emerge, exacerbating the risk of medical errors and decreased patient trust. High staff turnover can disrupt care continuity, and ethical dilemmas may lead to moral distress. Safety concerns, lapses in safety precautions, and unprofessional behavior pose additional risks to healthcare providers. Mitigating these risks necessitates healthcare organizations prioritizing staff well-being, offering stress management programs, mental health support, and fostering a supportive work environment, ultimately enhancing both patient care quality and healthcare professionals' overall health and job satisfaction. Hence, addressing these challenges is essential to support the well-being of medical professionals and ensure the provision of high-quality care.

Monitoring or measuring induced emotional charge is not only crucial to evaluate the quality of the work environment and to quantify its impact on each individual but also to plan mitigation measures and minimize deleterious effects. Quantifying negative emotions, the ones who should be mitigated, can be achieved through a wide range of methods. These include self-report scales and questionnaires, where providers complete assessments like the Perceived Stress Scale[1] or the Maslach Burnout Inventory[2] to gauge their stress levels subjectively. Physiological measures such as heart rate monitoring through wearable devices, cortisol level analysis in saliva, urine, or blood, skin conductance measurements, pupil dilation assessments, and EEG to record brain activity offer objective insights into stress responses. Behavioral observations, involving the analysis of speech patterns, facial expressions,[3] and body language, can provide additional indicators of stress levels. Biometric wearables,[4] activity and sleep trackers, thermal imaging, and short ecological momentary assessments (EMA) through mobile devices all contribute to real-time stress monitoring. By these methods, independently or combined, healthcare facilities can gain a comprehensive understanding of anxiety and stress levels among their staff and tailor support accordingly.

Speech analysis can be used to access a wide spectrum of diseases patterns[5-8] and it is also useful as an emotional biomarker, offering several advantages over other methods. First and foremost, it is a low-cost non-invasive and passive method, which means it does not require physical contact or instrumentation on the individual being assessed. This makes it comfortable and minimally disruptive during stress measurement, particularly in healthcare settings. Another significant advantage is its ability to provide objective measurements of stress-related features. By relying on computational analysis, it reduces the potential bias associated with self-report measures, enhancing the accuracy of stress assessment. Furthermore, speech analysis provides real-time monitoring, allowing for immediate intervention and support when stress is detected, which is particularly valuable in environments prone to heavy emotional load, like healthcare. It also allows for continuous assessment during conversations or interactions, offering a dynamic view of stress levels as they evolve over time. Additionally, machine learning algorithms can quantify emotion-related speech characteristics, enabling researchers and healthcare professionals to track and compare stress levels across individuals and situations. Lastly, speech analysis can be integrated with other biometric and physiological data sources, offering a more comprehensive understanding of stress responses.

In this manuscript, a stress estimation algorithm is proposed based on the classification of negative valence emotions in speech samples.

Material and Methods

The general pipeline for the proposed system is represented in Fig. 1. The description of each of the diagram's components will be detailed in next sub-sections.

Material

According with the multidimensional constructionist model of Lindquist[9] there are several studies that consider, as primary emotional dimensions, the valence (the pleasantness of a stimulus, happy to unhappy), the arousal (the intensity of emotion provoked by a stimulus, excited to calm), and the dominance (the degree of control exerted by a stimulus, from controlled to in control). For the purposes of the current study, the valence axis was defined as dominant and the emotions from the datasets, after being mapped on the emotional axis, were grouped as “negative” and “positive or neutral”.

The proposed system has an underlying machine learning decision model that must distinguish between two classes, negative emotions and neutral or positive emotions, making this a binary classification problem.

This statistical nature of such an approach requires the existence of annotated data that allows to perform a prior supervised learning stage, for adjusting the model's parameters. For this purpose, the RAVDESS dataset,[10] short for "Ryerson Audio-Visual Database of Emotional Speech and Song," was used. This is a popular multimodal dataset for emotion recognition research. It comprises a collection of audio and video recordings featuring 24 actors (12 males and 12 females) portraying a set of predefined emotions. It has metadata that encompasses seven emotional categories, with labels for calm, happy, sad, angry, fearful, disgust, and surprised, besides neutral. Each expression recording exists in two levels of emotional intensity (normal or strong), making a total of 1440 distinct recordings for the full dataset. The TESS dataset,[11] which stands for "Toronto Emotional Speech Set," was also used to enlarge the number of recordings.

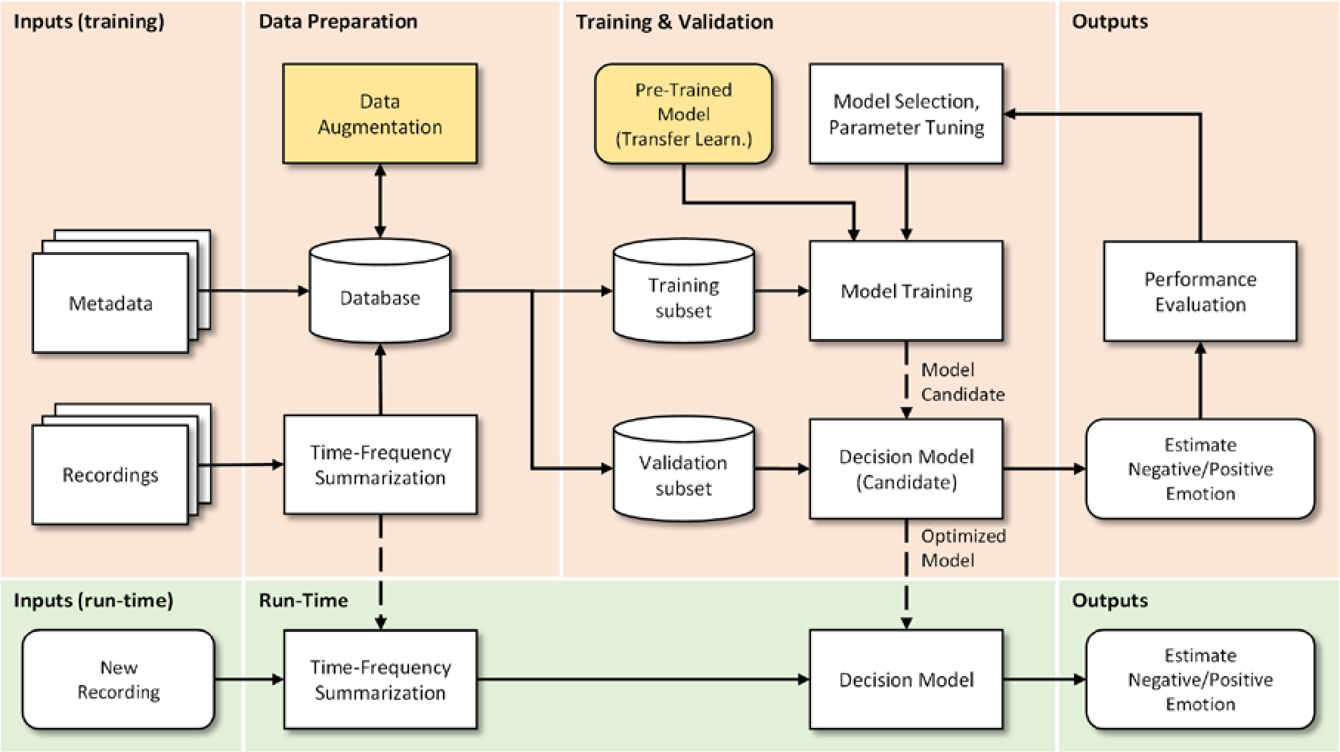


Figure 1: Functional diagram for the proposed emotional valence classification system, separated by training stage (top) and operation stage (bottom).

This dataset contains a collection of professionally acted, high-quality audio recordings of 200 different sentences spoken by two actresses. These sentences are designed to convey a range of similar set of emotional expressions. The dataset is composed of 2800 recordings.

The computational implementation of the proposed algorithms was performed using the Python programming language, the Librosa[12] package for audio processing, and PyTorch[13] packages.

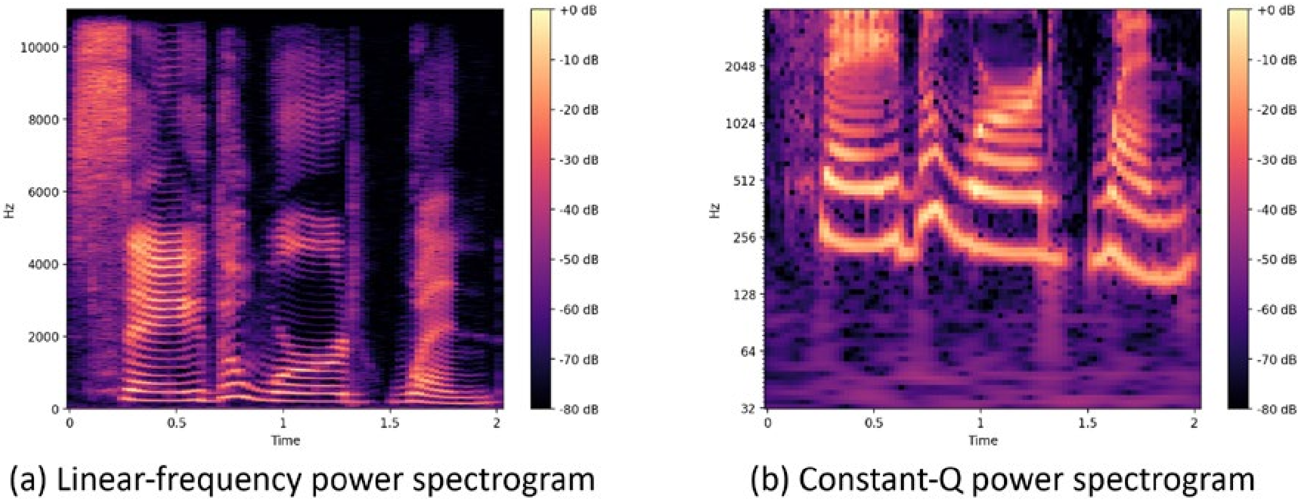
Methods

To further expand the number of recordings, the initial set was subjected to a data augmentation process where intensity was varied, and noise was added. For each recording, two recordings were generated, one with an increased intensity and another with a decreased intensity, both by a value of 10%. Additionally, for each recording, Gaussian noise was added, with random mean and standard deviation, while keeping intensity at 10% of two standard deviations. With this process, from the initial 4240 recordings, a new extended dataset of 16 960 recordings was obtained.

Due to computational performance limitations (both hardware and algorithms), earlier speech analysis systems used a set of features to represent, in a small dimension space, the most relevant characteristics. Praat[14] but also the Geneva Minimalistic Acoustic Parameter Set (GeMAPS)[15] or OpenSmile16 are popular toolkits for this purpose. However, the process of feature engineering (extraction and selection), requires specialized expertise and often involves extensive experimentation to find the optimal feature subset. In addition, the feature subset and the proceeding classification model, can be tightly connected, which introduces complexity in the exploration of the parameter set. Moreover, the feature selection process is susceptible to substantial information loss, potentially detrimental to system performance.

In the here proposed system, an end-to-end approach is used, eliminating the burden of the feature crafting step. For the time-frequency representation of the signal the constant Q power spectrogram was used. This is an alternative to the Fourier spectrogram or to the Mel Frequency Cepstral Coefficients (MFCCs). The constant-Q transform (CQT) features frequency bins that are logarithmically spaced and each frequency bin corresponds to a fixed fractional increase in pitch. In addition, by choosing an appropriate parameterization, it is possible to achieve good temporal resolution for transient events while still capturing fine-grained frequency details. CQT is also better in the representation of pitch, a highly relevant aspect of emotional speech.[17] In Fig. 2, a comparison of a linear Fourier spectrum can be observed side-by-side with a CQT power spectrum, for the same acoustic utterance. The recordings were processed in 2 seconds chunks with 0.5 seconds overlap.

For the decision model there is a wide range of network architectures. In this case, two distinct models were selected for evaluation, VGG-16[18] (with approx. 138 million parameters), for its good performance in speech related applications,[19] and SqueezeNet[20] (with approx. 1.25 million parameters), for



(a) Linear-frequency power spectrogram

(b) Constant-Q power spectrogram

Figure 2: Spectrograms representations of “Say the word when.” using a linear frequency spectrum (left) and a constant-Q spectrum (right).

its good compromise between computational footprint and modeling ability. Both networks require an input image with 224x224 pixels resolution.

For model training, a 10-fold validation strategy was used, where, for each iteration, the dataset was partitioned in 80% for training and 20% for validation. Using a transfer learning methodology, the models’ weight were initialized based on a pre-trained ImageNet[21] setup.

The training process was performed for 25 epochs and its duration was timed. Running a Python/PyTorch setup, in a i9-11900H with 32Gb RAM and a dedicated Nvidia GeForce RTX3050Ti, the VGG-16 model took around 3 hours to complete, while, for SqueezeNet, 25 minutes were necessary.

For system evaluation, the metrics precision, recall, F-score and confusion matrix were used as objective performance indicators. Considering TP as true positives, FP as false positives and FN as false negatives, the following equations were used:

$$\text{Precision} = TP / (TP+FP) \qquad (1)$$

$$\text{Recall} = TP / (TP+FN) \qquad (2)$$

$$\text{F-score} = 2*\text{Precision}*\text{Recall} / (\text{Precision}+\text{Recall}) \qquad (3)$$

Precision is a relevant metric that can be especially important in healthcare scenarios, where false positive errors can have serious consequences, however, it should not be considered in isolation. Recall, also known as sensitivity or true positive rate, can be a good complementary metric, since it focuses on the completeness of positive predictions, helping to understand how well the classifier captures all the positive instances in the dataset. Finally, F-score, is based on a trade-off between precision and recall, offering a more robust perspective of the results. However, on the downside, F-score is also prone to bias and it is invariant to TN, which should not be despised.

Results

As a first evaluation scenario, to assess the feasibility of the proposed solution, the classification of each emotion was observed independently. In Table 1, the confusion matrix that resulted from this classification task is presented, where, on each cell, the VGG-16 and SqueezeNet values are shown separated by a semicolon, respectively. The emotions that evidence the least amount of confusion are calm (81.8%), neutral (81.4%) and angry (81.1%), for VGG-16, and disgust (80%), angry (78.4%) and calm (77.9%), for SqueezeNet. For both decision models, it is possible to achieve a good discrimination for calm and angry emotions. Still in Table 1, in the two bottom rows, class related precision and recall values are presented. With VGG-16, the classification of happy (84.2%), fearful (83.8%) and sad (82.1%) achieved the best precision marks, and with SqueezeNet, the same set of emotions allowed the best performance, with the order fearful (80.9%), sad (80.6%) and happy (80.4%). The surprise label was the hardest to classify for both decision models. concerning recall, calm, Neutral and Angry presented values near 81%, using VGG-16 and the emotions calm, angry and disgust were classified around 2% lower, with SqueezeNet. Overall, SqueezeNet performed around 3.1% worse, on average, than VGG-16. F-score was 0.770 and 0.739 for VGG-16 and SqueezeNet, respectively.

For the second evaluation scenario, a binary classification task was considered, training the system to classify positive or negative valence for a given emotion. Using the previously described computational setup, a time of around 2 hours and 30 minutes was necessary to train the VGG-16 network while 25 minutes, as in the first scenarios, were required to train the SqueezeNet network.

The obtained results for the binary problem are shown in Table 2. For VGG-16, the precision (or positive predictive value-PPV) is 86.6% and the negative predictive value (NPV) is

Table 1: Confusion matrix for independent emotion classification, considering two distinct decision models, and class related evaluation metrics (precision and recall in the two bottom rows). Values are presented in percentage for models VGG-16 and SqueezeNet respectively, separated by a semicolon.

Real\Est.	Happy	Calm	Surprised	Neutral	Sad	Disgust	Fearful	Angry
Happy	67.6%; 63.4%	5.6%; 5.6%	5.6%; 5.6%	2.8%; 4.2%	5.6%; 5.6%	4.2%; 5.6%	2.8%; 4.2%	5.6%; 5.6%
Calm	2.6%; 2.6%	81.8%; 77.9%	3.9%; 5.2%	2.6%; 3.9%	2.6%; 3.9%	0.0%; 0.0%	2.6%; 2.6%	3.9%; 3.9%
Surprised	5.7%; 5.7%	0.0%; 2.9%	80.0%; 74.3%	0.0%; 0.0%	2.9%; 2.9%	5.7%; 5.7%	2.9%; 5.7%	2.9%; 2.9%
Neutral	0.0%; 0.0%	9.3%; 11.6%	7.0%; 7.0%	81.4%; 76.7%	0.0%; 0.0%	0.0%; 2.3%	2.3%; 2.3%	0.0%; 0.0%
Sad	1.4%; 1.4%	5.6%; 5.6%	0.0%; 1.4%	2.8%; 1.4%	77.5%; 76.1%	4.2%; 4.2%	4.2%; 4.2%	4.2%; 5.6%
Disgust	0.0%; 0.0%	0.0%; 0.0%	5.0%; 5.0%	5.0%; 5.0%	0.0%; 0.0%	80.0%; 80.0%	5.0%; 5.0%	5.0%; 5.0%
Fearful	3.8%; 5.0%	3.8%; 3.8%	5.0%; 5.0%	5.0%; 5.0%	2.5%; 2.5%	3.8%; 5.0%	71.3%; 68.8%	5.0%; 5.0%
Angry	1.4%; 2.7%	0.0%; 0.0%	2.7%; 2.7%	4.1%; 4.1%	4.1%; 4.1%	6.8%; 8.1%	0.0%; 0.0%	81.1%; 78.4%
Precision	84.2%; 80.4%	80.8%; 77.9%	60.9%; 56.5%	70.0%; 67.3%	82.1%; 80.6%	66.7%; 61.5%	83.8%; 80.9%	77.9%; 76.3%
Recall	67.6%; 63.4%	81.8%; 77.9%	80.0%; 74.3%	81.4%; 76.7%	77.5%; 76.1%	80.0%; 80.0%	71.3%; 68.8%	81.1%; 78.4%

90.0%, while SqueezeNet achieved 85.4% and 87.5%, respectively. In this scenario, all the values are above the best results obtained for the multiclass problem. For recall, a similar panorama is observed, with 88.5% (PPV) and 88.3% (NPV) for VGG-16, and 86.3% (PPV) and 87.5% (NPV). F-score was 0.884 and 0.866, for VGG-16 and SqueezeNet, both more than 0.1 above the previous results.

Discussion

The presented evaluation methodology and results were supported by pre-recorded datasets, which is useful for performance comparisons between distinct machine learning models.

The results that were obtained for the classification of each emotion independently allowed to understand the feasibility of the proposed processing pipeline and classification system. Performance could be further improved by exploring the parameter space or by fine-tuning the time-frequency representation. Nevertheless, the primary purpose of the system is to deliver estimates of positive or negative emotions, a more straightforward machine learning problem.

However, for real-life applications, further performance evaluation scenarios must be posed. Factors such as background noise, multiple voices or dialectal variations, among many others, can hamper acoustic pattern identification. Speech characteristics can also vary significantly based on the context of the conversation and the individual's communication style, creating an additional challenge.

The proposed emotion recognition technology can be applied to doctors and healthcare providers to enhance their well-being and the quality of patient care. By monitoring healthcare professionals' emotional states, such as stress and burnout, this technology can facilitate early intervention and support, ultimately benefiting their mental health. It can also aid in refining communication skills, promoting empathy, and providing valuable feedback for professional development. Furthermore, it enables healthcare providers to better understand and respond to patient emotions, fostering patient-centered care and improving overall patient satisfaction. Moreover, emotion recognition can assist in ethical decision-making, enhance teamwork and collaboration, and contribute to research and quality improvement efforts in healthcare.

Emotion recognition technology can be a valuable tool for patients as well, offering numerous benefits in healthcare settings. It can assist patients in understanding and managing their emotional well-being, particularly in chronic illness management, mental health care, or during acute medical situations. By providing real-time feedback on their emotional states, patients can become more aware of their mental health needs and can communicate more effectively with healthcare providers. Besides, emotion recognition systems can be used to assess pain levels objectively, aiding in pain management and ensuring that patients receive appropriate care and pain relief.

Finally, the integration of emotion recognition systems into telemedicine settings can help bridge the gap created

Table 2: Confusion matrix for positive and negative valence emotion classification. Values are presented in percentage for models VGG-16 and SqueezeNet respectively, separated by a semicolon.

Real\Est.	Positive	Negative
Positive	88.5% / 86.3%	11.5% / 13.7%
Negative	11.7% / 12.5%	88.3% / 87.5%
Precision	86.6%; 85.4%	90.0%; 87.5%
Recall	88.5%; 86.3%	88.3%; 87.5%

by remote consultations. Healthcare providers can assess patient emotions and well-being, compensating for the absence of in-person interaction. This process can be episodic or recurrent in time, tracking evolution or trends.

In all cases, consent and privacy must be paramount in the implementation of such technology, and either healthcare professionals and patients should have control over how their emotional data is used and shared.

Conclusion

In this paper an end-to-end emotional valence classification system based on speech samples was presented with the purpose of estimating emotional stressors. The presented developments were supported by widely known datasets and tools. Two decision models were trained for the purpose: VGG-16 and SqueezeNet. Remarkably, the SqueezeNet-based approach yielded comparable results to other reported works, all while maintaining a significantly smaller resource footprint. This feature enables deployment on mobile devices or low-performance computing platforms. The study demonstrates the viability of an end-to-end pipeline and hints at further enhancements that can be achieved by fine-tuning various parameters. When employed ethically and with the individuals' best interests in mind, emotion recognition technology can be a social empowering tool but also a resourceful indicator of healthcare. ■

Responsabilidades Éticas

Conflitos de Interesse: Os autores declaram a inexistência de conflitos de interesse na realização do presente trabalho.

Fontes de Financiamento: Este trabalho foi parcialmente financiado por Fundos Nacionais através da agência de financiamento portuguesa FCT - Fundação para a Ciência e a Tecnologia, no âmbito do projeto UIDB/50014/2020

Proveniência e Revisão por Pares: Não comissionado; revisão externa por pares.

Ethical Disclosures

Conflicts of Interest: The authors have no conflicts of interest to declare.

Financing Support: This work was partially financed by National Funds through the Portuguese funding agency FCT - Fundação para a Ciência e a Tecnologia, within the scope of the project UIDB/50014/2020

Provenance and Peer Review: Not commissioned; externally peer reviewed.

Correspondence / Correspondência:

Luís Pinto-Coelho – lfc@isep.ipp.pt,
Departamento de Física, Instituto Superior de Engenharia do Porto, Porto, Portugal
Rua Dr. Bernardino Ribeiro 431, 4249-051 Porto

Recebido / Received: 2023/10/06
Aceite / Accepted: 2023/10/12
Publicado / Published:2024/05/20

REFERENCES

1. Cohen S, Kamarck T, Mermelstein R. Perceived Stress Scale [Internet]. Chicago: APA Psyc Tests; 1983 [cited 2023 Sep 19]. Available from: https://psycnet.apa.org/doiLanding?doi: 10.1037%2Ft02889-000
2. Loera B, Converso D, Viotti S. Evaluating the Psychometric Properties of the Maslach Burnout Inventory-Human Services Survey (MBI-HSS) among Italian Nurses: How Many Factors Must a Researcher Consider? PLoS One. 2014;9:e114987. doi: 10.1371/journal.pone.0114987.
3. Coelho L, Reis S, Moreira C, Cardoso H, Sequeira M, Coelho R. Benchmarking Computer-Vision Based Facial Emotion Classification Algorithms While Wearing Surgical Masks. Engineering Proceedings. 2023 (in press).
4. Vieira FMP, Ferreira MA, Dias D, Cunha JPS. VitalSticker: A novel multimodal physiological wearable patch device for health monitoring. In: 2023 IEEE 7th Portuguese Meeting on Bioengineering (ENBENG). 2023. p. 100–3.
5. Deepa P, Khilar R. Speech technology in healthcare. Measurement. Sensors. 2022;24:100565.
6. Vigo I, Coelho L, Reis S. Speech- and Language-Based Classification of Alzheimer's Disease: A Systematic Review. Bioengineering. 2022;9:27.
7. Vieira H, Costa N, Sousa T, Reis S, Coelho L. Voice-Based classification of amyotrophic lateral sclerosis: where are we and where are we going? A systematic review. Neurodegener Dis. 2019;19:163-70. doi: 10.1159/000506259
8. Braga D, Madureira AM, Coelho L, Abraham A. Neurodegenerative Diseases Detection Through Voice Analysis. In: Abraham A, Muhuri PKr, Muda AK, Gandhi N, editors. Hybrid Intelligent Systems. Cham: Springer International Publishing; 2018. p. 213–23.
9. Lindquist KA. Emotions Emerge from More Basic Psychological Ingredients: A Modern Psychological Constructionist Model. Emotion Rev. 2013;5:356–68.
10. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One. 2018;13:e0196391. doi: 10.1371/journal.pone.0196391.
11. Pichora-Fuller MK, Dupuis K. Toronto emotional speech set (TESS) [Internet]. Borealis; 2020. [cited 2023 Sep 19].Available from: https://borealisdata.ca/citation?persistentId=doi:10.5683/SP2/E8H2MF

12. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, et al. librosa: Audio and Music Signal Analysis in Python. Proceedings of the 14th Python in Science Conference. 2015;18–24.
13. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library [Internet]. arXiv; 2019 [cited 2023 Sep 26]. Available from: http://arxiv.org/abs/1912.01703
14. Boersma P, Weenink D. Praat: doing phonetics by computer [Internet]. 2018. [cited 2023 Sep 19]. Available from: http://www.praat.org
15. Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Trans Affective Comput. 2016 Apr;7(2):190–202.
16. Eyben F, Wöllmer M, Schuller B. openSMILE -- The Munich Versatile and Fast Open-Source Audio Feature Extractor. MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. 2010. 1459 p.
17. Cabral JP, Oliveira LC. Emovoice: a system to generate emotions in speech. In: Interspeech 2006 [Internet]. ISCA; 2006 [cited 2023 Sep 26]. p. paper 1645-Wed2BuP.3-0. Available from: https://www.isca-speech.org/archive/interspeech_2006/cabral06_interspeech.html
18. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [Internet]. arXiv; 2015 [cited 2023 Sep 26]. Available from: http://arxiv.org/abs/1409.1556
19. de Lope J, Graña M. An ongoing review of speech emotion recognition. Neurocomputing. 2023;528:1–11.
20. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size [Internet]. arXiv; 2016 [cited 2023 Sep 26]. Available from: http://arxiv.org/abs/1602.07360
21. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition [Internet]. 2009 [cited 2023 Sep 26]. p. 248–55. [cited 2023 Sep 26] Available from: https://ieeexplore.ieee.org/document/5206848