Safety Management and Human Factors, Vol. 189, 2025, 182–188
https://doi.org/10.54941/ahfe1006584

Vocal Markers in Aviation of Workload, Stress, Fatigue, and Sleepiness: A Protocol Validation Study

Martina Gnerre and Federica Biassoni

Department of Psychology, Catholic University of the Sacred Heart, Largo Gemelli 1, Milan, Italy

ABSTRACT

This study is composed of two parts: the first part is a systematic review, and the second part is a protocol validation study. The systematic review aims to summarize and consolidate evidence from existing studies on the impact of workload, stress, fatigue, and sleepiness on speech, focusing on identifying specific vocal markers associated with these states within the context of aviation. Using PRISMA guidelines, we performed a comprehensive search of electronic databases, including Scopus, ScienceDirect, PsycINFO, and Web of Science. Twenty studies met the inclusion criteria and were analyzed to extract consistent vocal features indicative of these psychophysiological states in pilots and air traffic controllers (ATCs). Key findings from the review indicate that stress and workload are associated with increased vocal intensity and pitch, reflecting heightened sympathetic nervous system activation. Conversely, fatigue and sleepiness manifest through reduced vocal energy, slower speech rates, and increased pauses, indicative of diminished central nervous system activity. Mel-frequency cepstral coefficients (MFCCs) were highlighted as reliable and versatile indicators across all states. Building on the insights from the systematic review, the second part of the study focuses on validating an analysis protocol designed to detect and classify psychophysiological states in real-world aviation scenarios starting from vocal behavior. This protocol builds on the vocal markers identified in the review and applies structured acoustic analysis techniques using with Parselmouth, a Python interface to Praat (Jadoul et al., 2018). Real-world audio recordings were collected from pilots and ATCs. These scenarios included high-stress emergencies and routine operations. The recordings were processed to extract vocal features, including pitch, intensity, speech rate, pause duration, and MFCCs. Machine learning models were trained and tested on these features to classify the vocal data into categories of workload, stress, fatigue, and sleepiness. Although preliminary analyses are still underway, the current phase focuses on feature extraction and classification strategy development. Performance metrics will be assessed in future phases once model training is finalized. The integration of this validated protocol into aviation safety protocols may offer promising prospects for enhancing performance monitoring and risk mitigation. Real-time vocal monitoring systems could provide immediate feedback to pilots and ATCs, enabling timely interventions to address stress or fatigue before they compromise safety. Future work will focus on testing the system in operational settings and exploring the integration of vocal monitoring with existing cockpit technologies and ATC systems to support real-world implementation.

Keywords: Vocal markers, Acoustic analysis, Aviation safety, Pilots, ATCs

INTRODUCTION

Speech analysis has emerged as a valuable tool for assessing cognitive and emotional states in high-stakes operational environments, including aviation (Goguen & Linde, 1983; Greeley et al., 2013; Rakas et al., 2023). Pilots and air traffic controllers (ATCs) frequently operate under conditions of elevated workload, stress, fatigue, and sleepiness, all of which can significantly impact performance and safety (Nealley & Gawron, 2015). Given that voice serves as a primary mode of communication in aviation, vocal changes can provide critical insights into an individual's physiological and psychological state. Prior research has demonstrated that stress and cognitive overload can lead to increased pitch (F0), vocal intensity, and speech rate, while fatigue and sleepiness are often associated with lower vocal energy, reduced speech rate, and prolonged pauses (Chen et al., 2006; Huttunen et al., 2011). These findings have driven interest in using non-invasive vocal monitoring techniques to enhance real-time assessment of operator well-being in aviation settings (Van Puyvelde et al., 2018). Advancements in machine learning and speech processing technologies have facilitated the development of automated systems capable of classifying psychophysiological states based on vocal markers. Deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown promising results in accurately detecting cognitive and emotional states from speech data. However, methodological variations across studies, such as differences in recording conditions and feature extraction techniques, pose challenges for standardization and broader implementation. This study builds upon previous findings by systematically reviewing the current literature on vocal markers associated with workload, stress, fatigue, and sleepiness in aviation context.

A REVIEW

The review followed the PRISMA guidelines, conducting a comprehensive search across Scopus, PsycINFO, ScienceDirect, and Web of Science. Studies were included if they examined the impact of workload, stress, fatigue, or sleepiness on speech, employed acoustic analysis techniques, and involved pilots or ATCs. A total of 20 studies met the inclusion criteria and were analysed based on their methodological approaches, acoustic parameters, and classification outcomes.

Data extraction focused on key acoustic features such as fundamental frequency (F0), intensity, speech rate, pause duration, jitter, shimmer, and Mel-frequency cepstral coefficients (MFCCs). The studies varied in their methodological designs, with some conducted in real-world operational environments, while others relied on simulations or laboratory-controlled settings. Given the heterogeneity of data collection and analysis procedures, a narrative synthesis was employed rather than a meta-analysis. Findings from the review indicate that workload and stress are typically associated with increased F0 and vocal intensity, reflecting heightened activation of the sympathetic nervous system (Magnusdottir et al., 2022; Luig & Sontacchi, 2014). Speakers under high cognitive load also exhibit increased speech rate and reduced variability in pitch and amplitude (Alpert & Schneider, 1988; Huttunen et al., 2011). However, results regarding jitter and shimmer were inconsistent, suggesting that individual vocal responses to stress may vary. MFCCs emerged as robust indicators across different studies, demonstrating their potential utility in identifying workload and stress-related changes in speech. Conversely, fatigue and sleepiness were characterized by a decrease in vocal energy, slower speech rates, increased pause duration, and more frequent disfluencies such as hesitations and elongations. Spectral flattening and shifts in formant frequencies were also observed, indicating reduced articulatory precision. These vocal changes align with physiological models of reduced central nervous system activity during fatigue and sleep deprivation. Machine learning approaches were employed in several studies to classify psychophysiological states based on vocal features. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), demonstrated high classification accuracy, often outperforming traditional machine learning techniques like support vector machines (SVMs) and random forests. Notably, multimodal approaches integrating vocal features with physiological or contextual data yielded the most reliable performance in detecting fatigue and stress states. Figure 1 illustrates the primary findings of this systematic review, emphasizing the role of MFCC parameters as central indicators of workload, stress, fatigue, and sleepiness in aviation. A network analysis was conducted to examine the relationships between these psychophysiological states and specific acoustic features. The analysis, based on statistical associations from the reviewed studies, visualizes the interconnected nature of vocal markers rather than isolated effects. The results support the hypothesis that stress and workload, characterized by heightened sympathetic activation, correlate with increased vocal intensity and pitch, whereas fatigue and sleepiness are linked to lower vocal energy, slower articulation, and increased pauses. This visualization provides a comprehensive framework for understanding vocal monitoring as a non-invasive tool for assessing cognitive and physiological states in aviation.

PROTOCOL VALIDATION STUDY

This study follows the findings of the systematic review on vocal markers associated with workload, stress, fatigue, and sleepiness in aviation. The aim of this protocol is to analyze speech data from two large datasets (the ATCO2 and TARTAN Aviation datasets) to validate the vocal markers identified in the review and develop a classification system for detecting these psychophysiological states in ATCs and pilots. The protocol validation study was designed to assess the effectiveness of vocal markers in detecting workload, stress, fatigue, and sleepiness in aviation settings. Based on the systematic review findings, we developed a structured approach for data collection, feature extraction, and classification. At this stage, the analysis protocol is being finalized and focuses on defining optimal acoustic features for classification. Performance evaluation of machine learning models represents a key objective in the next phase of the research. Once models are trained and validated, standard performance metrics such as accuracy, precision, and recall will be applied to assess classification quality.

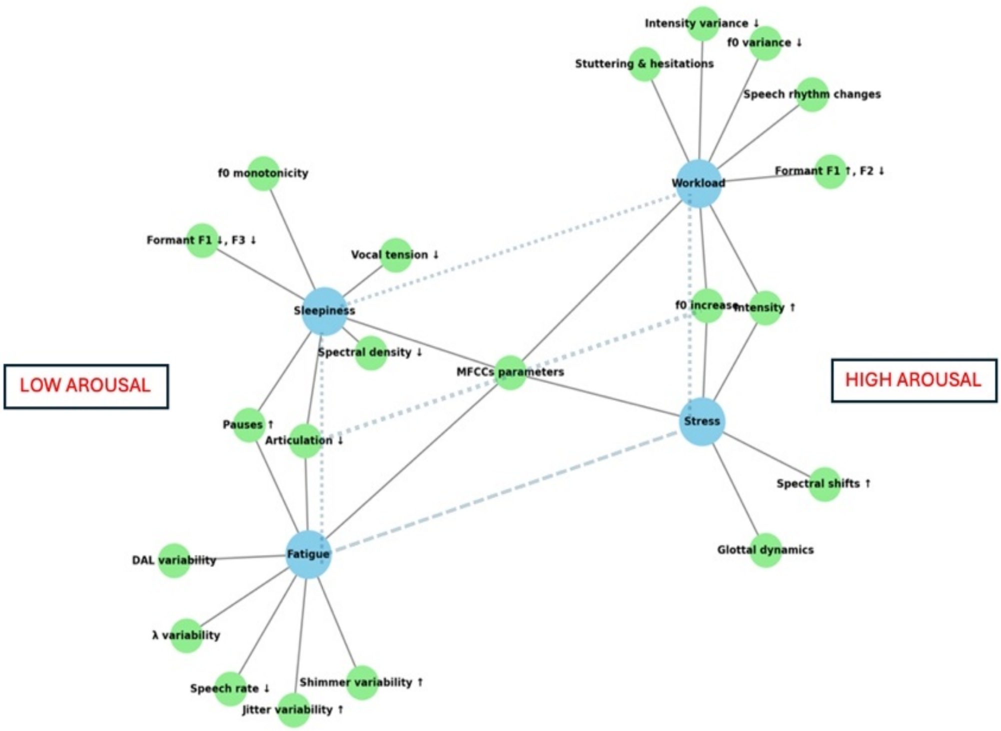

Figure 1: The diagram shows the connection between MFCC parameters and four states: workload, stress, fatigue, and sleepiness. Each state is linked to specific vocal changes, highlighting MFCCs' central role as versatile indicators for psychophysiological monitoring. The dashed lines represent a conceptual framework illustrating the complex interrelationships among fatigue, stress, workload, and sleepiness, emphasizing their mutual influence without a clear, unidirectional causal pathway.

Data Collection and Sources

Two large datasets are being used. The ATCO2 dataset (Zuluaga-Gomez et al., 2022) is serving as a large-scale corpus of real-world pilot–ATC communications, capturing a diverse range of operational scenarios. It encompasses both routine exchanges and high-stress emergency interactions, providing valuable linguistic and acoustic variability. Metadata includes timestamped communications, speaker roles (pilot or ATC), transmission clarity, and contextual annotations to facilitate structured analysis. The TARTAN Aviation dataset (Patrikar et al., 2024) is being developed as a multi-modal collection of real-world airport data, aiming to enhance research in aviation operations and automation. It integrates audio recordings of pilot–ATC communications, high-resolution image data, and aircraft trajectory information to offer a comprehensive view of air traffic environments. Structured to support studies on air traffic management, situational awareness, and human-machine interaction, the dataset is particularly relevant for advancing automation in aviation. Metadata includes communications, aircraft movement data, weather conditions, and operational context to aid in detailed analyses.

Speech Processing and Feature Extraction

Before analyzing vocal markers, the dataset is being cleaned and standardized to ensure data quality and comparability. This includes noise filtering to reduce background interference, speaker segmentation to distinguish between different voices (pilots and ATCs), and amplitude and pitch normalization to ensure uniform volume levels across recordings (following the procedure and formula outlined in Pell et al. (2009)). Once the audio is being preprocessed, a range of acoustic features is being extracted to capture different aspects of speech production and communication with Parselmouth, a Python interface to Praat (Jadoul et al., 2018) (see Table 1 for the main acoustic parameters).

Table 1: Acoustic parameters used and their associations with these psychophysiological states.

Category	Acoustic Parameter	Description	Associated Psychophysiological State(s)
Prosodic Features	Fundamental Frequency (F0)	Changes in pitch	Increased in stress & workload; decreased in fatigue & sleepiness
	Intensity	Loudness of speech	Increased in stress & workload; decreased in fatigue & sleepiness
	Speech Rate	Speed of spoken words	Increased in stress & workload; decreased in fatigue & sleepiness
Temporal Features	Pause Duration	Length of pauses between speech segments	Increased in fatigue & sleepiness
	Speech Continuity	Flow of speech without interruptions	Reduced in fatigue & sleepiness
Spectral Features	MFCCs	Mel-Frequency Cepstral Coefficients, capturing spectral shape	Reliable indicator across all states
	Formant Frequencies (F1-F3)	Resonance frequencies of speech	Altered in stress & fatigue
	Spectral Tilt	Energy distribution in frequency bands	Flattened in fatigue & sleepiness
Voice Quality Features	Jitter	Frequency perturbation	Increased in workload; inconsistent in stress
	Shimmer	Amplitude perturbation	Increased in workload & fatigue
	HNR (Harmonics-to-Noise Ratio)	Ratio of periodic to aperiodic energy	Lower in fatigue & sleepiness

Continued

Table 1: Continued

Category	Acoustic Parameter	Description	Associated Psychophysiological State(s)
Articulatory Features	Articulation Rate	Number of syllables per second	Reduced in fatigue & sleepiness
	Spectral Center of Gravity	Weighted average frequency in a speech signal	Shifted under stress & workload

Experimental Protocol and Validation Strategy

To ensure the reliability and validity of the dataset, speech samples are being independently classified by expert raters based on both operational conditions (e.g., routine, emergency) and physiological states (e.g., stress, fatigue). The classification process is being conducted by a panel of psychologists specialized in human factors and aviation experts with direct operational experience. Prior to annotation, raters are undergoing a structured training phase to enhance inter-rater reliability and minimize subjective bias. Annotations are performed using Praat TextGrid files, allowing for precise segmentation and labeling of speech events (Boersma & Van Heuven, 2001). The consistency of their classifications is being evaluated using inter-rater reliability measures, ensuring a robust and reproducible labeling process. Following classification, speech samples are undergoing acoustic analysis using Parselmouth (Jadoul et al., 2018). A refinement procedure is then being applied to optimize feature selection. Specifically, a correlation-based approach is being employed to identify and remove redundant or acoustically non-informative features, retaining only those parameters that provide meaningful differentiation of workload and physiological states. This process is enhancing the sensitivity and specificity of speech-based assessments by ensuring that only the most diagnostically relevant acoustic markers are being preserved.

Expected Contributions

This protocol establishes a standardized framework for acoustic analysis in aviation communication, allowing for an objective and precise assessment of workload, stress, fatigue, and sleepiness through speech characteristics. The findings will support the development of non-invasive monitoring tools aimed at improving aviation safety and operator well-being. By integrating real-world and controlled speech datasets, this study lays the groundwork for a scalable, real-time vocal monitoring system tailored for aviation safety applications. Future developments will involve real-time implementation tests in simulated and operational contexts, assessing usability and accuracy under live conditions. In addition, the protocol will be adapted to integrate with existing cockpit and ATC interfaces to support seamless deployment in current aviation infrastructures.

Ethical Considerations

All data used in this study are derived from publicly available datasets, ensuring compliance with ethical guidelines. The analysis does not include personally identifiable information, ensuring full anonymity of recorded speech.

REFERENCES

Alpert, M., & Schneider, S. J. (1988). Voice-stress measure of mental workload. *NASA. Langley Research Center, Mental-State Estimation, 1987*.

Boersma, P., & Van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International*, *5*(9/10), 341–347

Chen, F. (2006). Attention, Workload and Stress. *Designing Human Interface in Speech Technology*, 53–94.

Goguen, J. A., & Linde, C. (1983). Linguistic methodology for the analysis of aviation accidents (No. NASA-CR-3741).

Greeley, H. P., Roma, P. G., Mallis, M. M., Hursh, S. R., Mead, A. M., & Nesthus, T. E. (2013). Field study evaluation of cepstrum coefficient speech analysis for fatigue in aviation cabin crew (No. DOT/FAA/AM-13/19). United States. Office of Aerospace Medicine.

Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., & Leino, T. (2011). Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. Applied ergonomics, 42(2), 348–357.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, 71, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001

Luig, J., & Sontacchi, A. (2014). A speech database for stress monitoring in the cockpit. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, *228*(2), 284–296.

Magnusdottir, E. H. (2022). Cognitive workload and voice. Psychological Reports, 128(6), 3150–3162.

Nealley, M. A., & Gawron, V. J. (2015). The effect of fatigue on air traffic controllers. The International Journal of Aviation Psychology, 25(1), 14–47.

Patrikar, J., Dantas, J., Moon, B., Hamidi, M., Ghosh, S., Keetha, N.,... & Scherer, S. (2024). TartanAviation: Image, Speech, and ADS-B Trajectory Datasets for Terminal Airspace Operations. *arXiv preprint arXiv:2403.03372*.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. Journal of Phonetics, 37, 417–435.

Rakas, J., Sohn, S., Keslerwest, L., & Krozel, J. (2023). Deep Speech Pattern Analysis of Controller-Pilot Voice Communications for Enhancing Future Aviation Systems Safety. In AIAA AVIATION 2023 Forum (p. 4410).

Rakas, J., Vallioor, V. K., Krozel, J., Kostiuk, P. F., & Mohen, M. T. (2024). Controller-Pilot Voice Communication and Intent Monitoring for Future Aviation Systems Safety. In *AIAA AVIATION FORUM AND ASCEND 2024* (p. 3942).

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in psychology*, *9*, 1994.

Zuluaga-Gomez, J., Veselý, K., Szöke, I., Blatt, A., Motlicek, P., Kocour, M.,... & Klakow, D. (2022). Atco2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications. arXiv preprint arXiv:2211.04054.