AI Schizophrenia Diagnosis through Speech Features F0 and MFCC

Felipe Lage Teixeira[1,2,4,6], Joana Fernandes[1,7], Adriana Ondina Pestana Santos[8], J.L Pio Abreu[9], Salviano Pinto Soares[4,5,6], and João Paulo Teixeira[1,2,3]

[1] Research Centre in Digitalization and Intelligent Robotics (CEDRI), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal.
[felipe.lage;joana.fernandes; joaopt]@ipb.pt,
home page: https://www.cedri.ipb.pt

[2] Applied Management Research Unit (UNIAG), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal.
uniag@ipb.pt,
home page: https://uniag.ipb.pt/index.php

[3] Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal.
sustec@ipb.pt,
home page: https://sustec.ipb.pt

[4] Engineering Department, School of Sciences and Technology, University of Trás-os-Montes and Alto Douro (UTAD), Quinta de Prados, 5000-801 Vila Real, Portugal.
felipe.lage@ipb.pt; salblues@utad.pt,
home page: https://www.utad.pt/ect/

[5] Intelligent Systems Associate Laboratory (LASI), University of Aveiro, 3810-193 Aveiro, Portugal.
sec@lasi-research.pt,
home page: https://lasi-research.pt/

[6] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, 3810-193 Aveiro, Portugal.
rtsaude@ieeta.pt,
home page: https://www.ieeta.pt/

[7] Faculty of Engineering of Porto (FEUP),4200-465 Porto, Portugal.
feup@fe.up.pt,
home page: https:www.fe.up.pt

[8] Instituto Português de Oncologia de Coimbra Francisco Gentil Martins EPE, 3000-075 Coimbra, Portugal
adriana538santos@gmail.com,
home page: https://www.ipocoimbra.min-saude.pt/

[9] Faculty of Medicine of the University of Coimbra, 3000-548 Coimbra, Portugal
pioabreu@icloud.com,
home page: https://www.uc.pt/fmuc/

Abstract. Schizophrenia affects over 20 million people globally and is often undetected in its early stages. Speech has unique characteristics

that can help identify mental illnesses, including schizophrenia, which usually manifests through slower, repetitive, or incoherent speech patterns. By extracting acoustic features like fundamental frequency (F0) and Mel Frequency Cepstral Coefficients (MFCCs) and applying machine learning, we can identify patterns that distinguish healthy individuals from those with schizophrenia. In this work, was achieved 95% accuracy to classify between schizophrenic and healthy people through speech.

Keywords: Schizophrenia,, Ensemble Bagged Trees, Ensemble Boosted Trees, Narrow Neural Network

1 Introduction

Mental health is a dynamic state of internal balance that enables individuals to use their abilities harmoniously, in alignment with the values of the surrounding society. This balance is sustained through the interaction of cognitive knowledge, social skills, and the ability to regulate, express, and recognize one's own emotions. Thus, adverse life events, social functions, and the interaction between body and mind can directly influence mental health [1, 2].

People with severe mental illnesses live, on average, 10 to 20 years less than the general population [3], especially when the illness becomes chronic. Schizophrenia, a chronic and severe mental illness with heterogeneous manifestations, directly affects the lifestyle of over 20 million adults worldwide [4]. According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), schizophrenia is characterized by abnormalities in one or more of the following areas: delusions, hallucinations, disorganized thinking (speech), severely disorganized or abnormal motor behavior (including catatonia), and negative symptoms [5].

Early diagnosis of these conditions is crucial to ensure proper treatment and patient recovery. However, current diagnostic methods, which rely on observations and interviews conducted by psychiatrists, may be subject to human error and can be time-consuming. Cognitive system assessment through features such as attention, perception, memory (working and explicit), cognitive control, and speech behavior offers a way to characterize mental disorders. Speech and language, in turn, provide valuable insights into human thought, allowing for the analysis of aspects such as semantic and emotional content, semantic coherence, syntactic structure, and complexity [6].

Speech is a complex and variable signal that carries both linguistic and emotional information. Cognitive and thought disorders manifest in the way speech is produced and in the content of what is said. Linguistic disorganization, for example, can be related to neurological dysfunctions, such as in epilepsy, when the brain's non-dominant hemisphere is affected [7, 8, 6].

Technologies applied to mental health must consider ethical and responsibility issues, respecting patient privacy. Implementing AI (Artificial Intelligence) in psychiatry can be challenging due to disagreements between doctors and researchers. Diagnostic support tools do not replace medical diagnoses, which re-

main the responsibility of professionals. In the case of schizophrenia, current methods rely on observations and interviews, using several scales such as Positive and Negative Syndrome Scale (PANSS) to assess disease severity [6].

Although widely accepted by the medical community, these schizophrenia assessment scales have limitations, such as subjectivity, which can lead to inconsistent evaluations among professionals. The complexity and variability of symptoms make standardization difficult, compromising the accuracy of measurements. Factors like cultural and linguistic barriers, lack of standardization across scales, the time required for assessment, and incomplete symptom coverage also affect the reliability of these scales. Combining these scales with additional clinical features, such as speech parameters, can help mitigate these disadvantages. The use of these features provides greater objectivity, sensitivity, and ease of application, improving diagnostic support. AI based on biological characteristics can increase accuracy, reduce the time and resources needed for diagnosis, and help prevent psychotic episodes.

In this study, MFCC and fundamental frequency (F0) parameters were applied to various machine learning tools to identify the most promising one and determine whether combining these different features improves the accuracy of identifying speech with or without a diagnosis of schizophrenia.

This article is organized into five chapters. The current chapter serves as an introduction, setting the stage for the subject under study. Chapter 2 provides a review of the state of the art, highlighting similar works and key findings from other authors. Chapter 3 details the methodology used, offering insight into the approach taken. In Chapter 4, the results are presented alongside an discussion of their implications. Finally, Chapter 5 draws together the conclusions reached through this research.

2 State of the Art

Technological advances have enabled the application of AI over time. The use of signal processing techniques offers the advantage of capturing details that are imperceptible to the naked eye, making parameter extraction a crucial part of this process [9, 10]. Although the literature includes studies that examine more parameters than those addressed here [11, 12], this work focuses on two: fundamental frequency and Mel-frequency cepstral coefficients (MFCC). These are two of the main acoustic features that allow the detection of speech alterations. As mentioned in the introduction, schizophrenia presents speech abnormalities as one of its symptoms.

Gosztolya et al. [13] used only temporal features extracted from spontaneous speech, such as articulation rate, speech tempo, utterance duration, number of pauses, pause duration, pause frequency, and average pause duration. The authors achieved 70–80% accuracy in classifying individuals with a schizophrenia diagnosis [13]. Other studies used two categories of features. Kliper et al. [14] combined temporal and prosodic features to identify schizophrenia, depression, and control subjects. The parameters analyzed included spoken ratio, utterance

duration, gap duration, pitch range, pitch standard deviation, power standard deviation, mean waveform correlation, mean jitter, and mean shimmer. These parameters allowed the classification of control vs. schizophrenia with 76.19% accuracy, control vs. depression with 87.5% accuracy, and schizophrenia vs. depression with 71.43% accuracy. For multiclass classification, they achieved 69.77%. Martínez-Sánchez et al. [15] and Rapcan et al. [16] showed that patients with schizophrenia tend to have slower speech, reduced pitch variability, and a higher number of pauses. Rapcan et al. [16] investigated fundamental frequency (F0) and the relative variation of vocal pitch, and using temporal and prosodic features, attempted to study total speech duration but did not find statistical significance. They argued that the lack of academic qualifications of the analyzed subjects compromised the results.

Compton et al. [17] also used two categories of features, in this case, prosodic and spectral. They studied schizophrenic patients and healthy subjects with and without aprosody. They concluded that patients with aprosody present lower F0, F2, and intensity/loudness values. Some authors used features from all four categories. Agurto et al. [18] were able to predict psychosis with 90% accuracy using prosodic, spectral, temporal, and statistical measures. Speech analysis was also combined with other parameters. In Chakraborty et al. [19], the algorithm's performance improved when body movements were included as input parameters. For example, [19] applied low-level descriptors (LLD) and body movements to detect negative symptoms. The LLD set includes intensity, loudness, MFCC (12), pitch (F0), probability of voicing, F0 envelope, 8 LSF (Line Spectral Frequencies), and Zero-Crossing Rate. Using an SVM classifier with only the LLD features, the authors obtained an accuracy of 79.49%. When these features were combined with body movements, the accuracy improved to 86.36%.

3 Methodology

In this work, a diarization process was required (It corresponds to the identification of the speaker and the time they speak, in other words, who is speaking and when). The speech segments corresponding to healthy subjects and those diagnosed with schizophrenia were manually separated. A total of 3,632 speech excerpts were used, with all participants in the recordings being native European Portuguese speakers. These excerpts come from 19 subjects diagnosed with schizophrenia and 54 healthy subjects [20, 21]. The data was divided into two sets: training (85%) with 3088 observations, and testing (15%). Parameters were extracted using 40 ms frames with a 20 ms overlap, utilizing a *periodic Hamming window* [22]. Cross-validation was employed (to avoid overfitting), splitting the data into 5 subsets (folds=5).

For all audio samples, the sampling rate was checked, and resampling to 44.1 kHz was performed when necessary. This step was taken because some recordings were obtained in different environments and using different equipment, as mentioned in [20]. The extraction of F0 was performed using the *Normalized Correlation Function - NCF* [23]. Thirteen MFCC parameters were extracted

from each frame. Considering the longest excerpt (75.48 seconds, healthy subject, male, 25 years old), the maximum number of F0 parameter samples was 3773, MFCC samples 49049, and 52822 for the combined MFCC and F0 set. Combinations of features, F0 only, MFCC only, and a combination of F0 and MFCC were trialed.
Ensemble Boosted Trees combines multiple decision trees sequentially, correcting errors from each previous iteration to enhance accuracy, making it ideal for complex problems, though it may suffer from overfitting. Ensemble Bagged Trees also combines trees but uses "bagging," where each tree is trained on random samples, increasing robustness and reducing model variance, making it effective on noisy data. A Narrow Neural Network is a neural network with few layers and neurons, suitable for smaller-scale problems as it requires less memory and reduces the risk of overfitting.

In studies of this kind, it is common to test various tools. For this purpose, MATLAB®2023a was used, specifically the *Classification Learner tool*, where all available model options were tested. However, tools such as *Naive Bayes and Quadratic Discriminant* were discarded, as they were found to be less promising according to the study in [6].

4 Results and Discussion

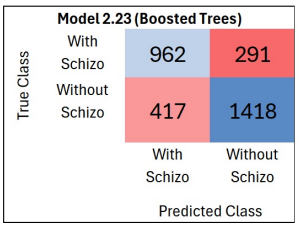
In this study, F0 and MFCC features were applied because the literature has shown that they can indicate changes in individuals' speech. On the other hand, the aim was to maintain state-of-the-art accuracy while reducing computational load to ensure efficient processing, making it feasible for real-time applications (a future goal).

For the prediction made using F0, the most promising result was found with the ensemble boosted trees tool, achieving an accuracy of 77% on the validation set (figure 1a) and 72.98% on the test set (see table: 1), as shown in the confusion matrix in figure 2a.

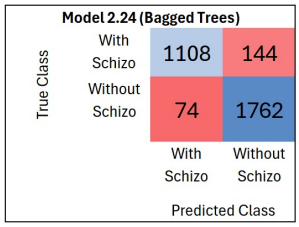
For the prediction made using MFCCs, the most promising result was found with the ensemble bagged trees tool, achieving an accuracy of 92.94% on the validation set (figure 1b) and 93.38% on the test set (see table: 1), as shown in the confusion matrix in figure 2b. This model had an average prediction speed of 0.81 obs/sec. Training time was 22,586 seconds (Processor: Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz, RAM: 16 GB).

For the prediction made using MFCCs and F0 simultaneously, the most promising result was achieved with the narrow neural network tool (fully connected layer: 1, first layer size: 10, activation: ReLU, iteration limit: 1000), reaching an accuracy of 92.13% on the validation set (figure 1c) and 95% on the test set (see table: 1), as shown in the confusion matrix in figure 2c. This model had an average prediction speed of 1.4 obs/sec. Training time was 12,031 seconds (Processor: AMD EPYC 7351 16-Core Processor @ 2.40 GHz, RAM: 64 GB).

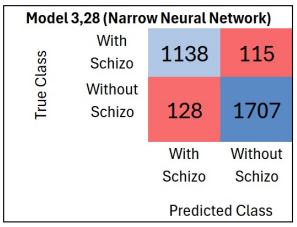
In figure 3 "Com esquizofrenia" means *With Schizophrenia* and "Sem esquizofrenia" means *Without Schizophrenia*.



(a) Confusion matrix of validation set with F0 features.

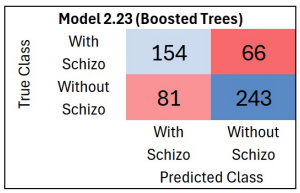


(b) Confusion matrix of validation set with MFCC's features.

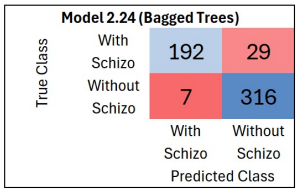


(c) Confusion matrix of validation set with F0 and MFCC's features.

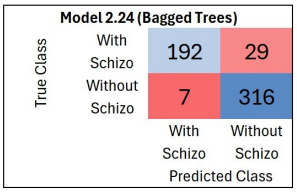
Fig. 1: Validation confusion matrix for F0, MFCC's and both.



(a) Confusion matrix of test set with F0 features.



(b) Confusion matrix of test set with MFCC's features.



(c) Confusion matrix of test set with F0 and MFCC's features.

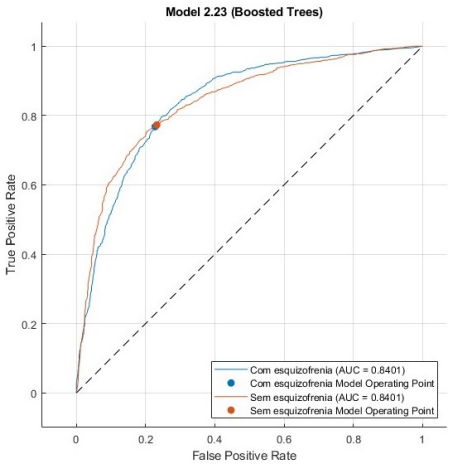
Fig. 2: Test confusion matrix for F0, MFCC's, and both.

Table 1: Best obtained results with F0, MFCC's and MFCC's+F0.

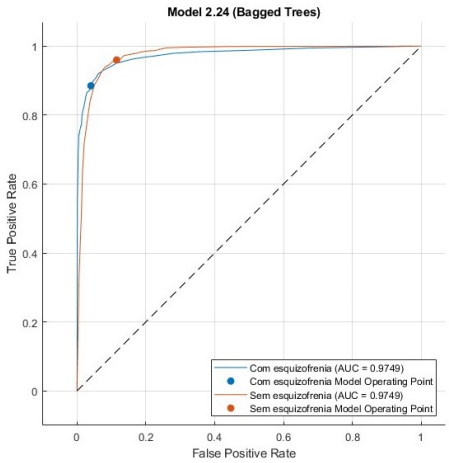
Features	**Model Name**	**Validation Acc. (%)**	**Test Acc. (%)**
F0	Ensemble Boosted Trees	77,07	72,98
MFCC's	Ensemble Bagged trees	92,94	93,38
MFCC's + F0	Narrow Neural Network	92,13	95,04

Figure 3 shows that all models have an equal AUC (Area Under the Curve) between the two categories analyzed, indicating a similar overall ability to distinguish between the classes. However, the operating points of the models and the curvature of the ROC differ. The closer the operating point is to the top-left corner of the graph, the higher the true positive rate (TPR) and the lower the false positive rate (FPR).

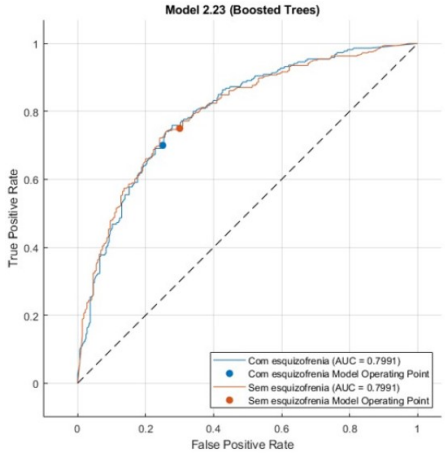
For example, in Figure 3b, with an AUC of 0.9749, the model separates the classes with an accuracy probability of 97.49%, while in Figure 3a this probability is approximately 84%. This pattern is also observed in the ROC curves of the test sets, where Figure 3d shows greater reliability than the model represented in Figure 3c. Comparing the model trained with F0, it can be seen that its AUC



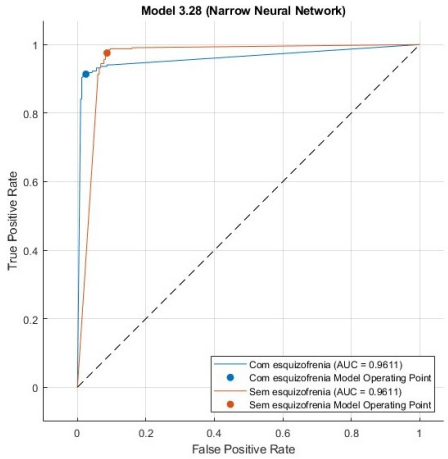
(a) Validation ROC curve with F0 features.



(b) Validation ROC curve with MFCC's.



(c) Test ROC curve with F0 features.



(d) Test ROC curve with MFCC's and F0 features.

Fig. 3: Receiver Operating Characteristic (ROC) curve of the best and worst results obtained in the validation and test set.

was higher in the validation set (0.8401, Figure 3a) and decreased by about 4% in the test set (0.7991, Figure 3c).

The implementation of additional features, such as GTCC (gammatone cepstral coefficients), Zero Cross Rating, harmonic ratio, and short-time energy, was tested, but the results worsened considerably by around 13% when compared to our best results. This suggests that adding multiple features does not always enhance the predictive power of the model.

Table 2 shows other results achieved by applying the parameters to other tools tested.

Table 2: Other obtained results with F0, MFCC's and MFCC's+F0.

Features	Tools	Validation acc(%)	Test acc(%)
F0	Tree	72,28	69,85
	SVM	71,18	67,10
	KNN	71,50	70,59
	Ensemble	77,07	72,98
	Neural Network	66,87	63,24
	Kernel	73,90	69,30
MFCC	Tree	89,83	90,81
	SVM	90,71	92,65
	Ensemble	92,94	93,38
	Neural Network	91,87	92,83
MFCC + F0	Tree	89,93	87,13
	SVM	91,09	91,18
	KNN	82,74	81,99
	Ensemble	93,43	92,10
	Neural Network	92,13	95,04

5 Conclusions

Studies and tests so far show that the use of technology in healthcare has increased over time. However, there are still some limitations in mental health due to its sensitive nature and the ethical and bureaucratic challenges it involves.

These types of systems can be valuable in assisting medical decisions. While there may be some clinical resistance to adopting such tools, scientific evidence shows that the parameters analyzed in this work are directly linked to speech patterns in individuals with schizophrenia. This study focused specifically on detecting schizophrenia with a maximum accuracy of 95% with both features, when compared to the state of the art (best result of 86.36% accuracy) shows an improvement of 8.64%.

Future work will focus on integrating emotional state detection with schizophrenia diagnosis through speech, despite the challenge of lacking standardized emotional labels for patients with schizophrenia. Additionally, the system's interface will be refined and tested in clinical settings, enabling feedback from medical specialists and potential integration of biometric data, such as facial analysis and movement patterns.

Acknowledgements. This work was supported by national funds through FCT/ MCTES (PIDDAC): CeDRI, UIDB/05757/2020 (DOI: 10.54499 /UIDB/05757/ 2020) and UIDP/05757 /2020 (DOI: 10.54499/ UIDP /05757/ 2020); and SusTEC, LA/P/ 0007/2020 (DOI: 10.54499/ LA/P/0007/2020).

The authors are grateful to Dr. Adriana Pestana-Santos for the access provided to the expert-annotated dataset of patients diagnosed with schizophrenia.

References

1. Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116, 2020.
2. Silvana Galderisi, Andreas Heinz, Marianne Kastrup, Julian Beezhold, and Norman Sartorius. A proposed new definition of mental health. *Psychiatria Polska*, 51(3):407–411, 2017.
3. Sam Manger. Lifestyle interventions for mental health. *Australian Journal for General Practitioners*, 48(10):670–673, October 2019.
4. Fiona J Charlson, Alize J Ferrari, Damian F Santomauro, Sandra Diminic, Emily Stockings, James G Scott, John J McGrath, and Harvey A Whiteford. Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophrenia Bulletin*, 44(6):1195–1203, 05 2018.
5. A. Barbato. Schizophrenia and public health. https://apps.who.int/iris/bitstream/handle/10665/63837/WHO_MSA_NAM_97.6.pdf?sequence=1, 1997. Accessed: 25 September 2024.
6. Felipe Lage Teixeira, Miguel Rocha e. Costa, José Pio Abreu, Manuel Cabral, Salviano Pinto Soares, and João Paulo Teixeira. A Narrative Review of Speech and EEG Features for Schizophrenia Detection: Progress and Challenges. *Bioengineering*, 10(4):1–31, 2023.
7. S. C. Park, K. Kim, O. J. Jang, H. J. Yoon, S. H. Jang, S. W. Kim, B. J. Lee, J. H. Park, K. U. Lee, and J. Choi. Network analysis of language disorganization in patients with schizophrenia. *Yonsei Medical Journal*, 61(8):726–730, August 2020.
8. A. S. Peters, J. Rémi, C. Vollmar, J. A. Gonzalez-Victores, J. P. S. Cunha, and S. Noachtar. Dysprosody during epileptic seizures lateralizes to the nondominant hemisphere. *Neurology*, 77:1482–1486, 2011. [CrossRef] [PubMed].
9. João Paulo Teixeira, Diamantino Freitas, Daniela Braga, Maria João Barros, and Vagner Latsch. Phonetic events from the labeling the european portuguese database for speech synthesis, feup/ipb-db. page 1707 – 1710, 2001.
10. João Paulo Teixeira and Diamantino Freitas. Segmental durations predicted with a neural network. page 169 – 172, 2003.
11. João Paulo Teixeira., Joana Fernandes., Filipe Teixeira., and Paula Odete Fernandes. Acoustic analysis of chronic laryngitis. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - BIOSIGNALS*, pages 168–175. INSTICC, SciTePress, 2018.
12. Joana Fernandes, Leticia Silva, Felipe Teixeira, Victor Guedes, Juliana Santos, and João P. Teixeira. Parameters for vocal acoustic analysis - cured database. *Procedia Computer Science*, 164:654–661, 2019. CENTERIS 2019 - International Conference on ENTERprise Information Systems / ProjMAN 2019 - International Conference on Project MANagement / HCist 2019 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019.
13. Gábor Gosztolya, Anita Bagi, Szilvia Szalóki, István Szendi, and Ildikó Hoffmann. Identifying schizophrenia based on temporal parameters in spontaneous speech. In *Interspeech 2018*, pages 3408–3412, 2018.

14. Roi Kliper, Shirley Portuguese, and Daphna Weinshall. Prosodic analysis of speech and the underlying mental state. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, 2015.
15. Francisco Martínez-Sánchez, José Antonio Muela-Martínez, Pedro Cortés-Soto, Juan José García Meilán, Juan Antonio Vera Ferrándiz, Amaro Egea Caparrós, and Isabel María Pujante Valverde. Can the acoustic analysis of expressive prosody discriminate schizophrenia? *The Spanish Journal of Psychology*, 18:E86, 2015.
16. Viliam Rapcan, Shona D'Arcy, Sherlyn Yeap, Natasha Afzal, Jogin Thakore, and Richard B. Reilly. Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical Engineering Physics*, 32(9):1074–1079, 2010.
17. Michael T. Compton, Anya Lunden, Sean D. Cleary, Luca Pauselli, Yazeed Alolayan, Brooke Halpern, Beth Broussard, Anthony Crisafio, Leslie Capulong, Pierfrancesco Maria Balducci, Francesco Bernardini, and Michael A. Covington. The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech. *Schizophrenia Research*, 197:392–399, 2018.
18. Carla Agurto, Mary Pietrowicz, Raquel Norel, Elif K. Eyigoz, Emma Stanislawski, Guillermo Cecchi, and Cheryl Corcoran. Analyzing acoustic and prosodic fluctuations in free speech to predict psychosis onset in high-risk youths. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5575–5579, 2020.
19. Debsubhra Chakraborty, Zixu Yang, Yasir Tahir, Tomasz Maszczyk, Justin Dauwels, Nadia Thalmann, Jianmin Zheng, Yogeswary Maniam, Nur Amirah, Bhing Leet Tan, and Jimmy Lee. Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6024–6028, 2018.
20. Adriana Pestana-Santos, Luís Loureiro, Vítor Santos, and Irene Carvalho. Patients with schizophrenia assessing psychiatrists' communication skills. *Psychiatry Research*, 269(August):13–20, 2018.
21. Rui Pedro Lopes, Bárbara Barroso, Leonel Deusdado, André Novo, Manuel Guimarães, João Paulo Teixeira, and Paulo Leitão. Digital technologies for innovative mental health rehabilitation. *Electronics (Switzerland)*, 10(18):1–15, 2021.
22. Felipe L. Teixeira, Salviano Pinto Soares, J.L. Pio Abreu, Paulo M. Oliveira, and João P. Teixeira. Comparative Analysis of Windows for Speech Emotion Recognition Using CNN. pages 233–248, 2024.
23. B. S. Atal. Automatic Speaker Recognition Based on Pitch Contours. *The Journal of the Acoustical Society of America*, $45(1_Supplement) : 309 - -309, 1969$.