

Comparative Evaluation of Acoustic Feature Extraction Tools for Clinical Speech Analysis

Anna Seo Gyeong Choi¹, Alexander Richardson², Ryan Partlan², Sunny Tang^{*2}, Sunghye Cho^{*3}

¹Information Science, Cornell University, USA

²Department of Psychiatry, Northwell Health, USA

³Linguistic Data Consortium, Department of Linguistics, University of Pennsylvania, USA

sc2359@cornell.edu, arichardson7@northwell.edu, rpartlan@northwell.edu,
stang3@northwell.edu, csunghye@ldc.upenn.edu

Abstract

This study compares three acoustic feature extraction toolkits—OpenSMILE, Praat, and Librosa—applied to clinical speech data from individuals with schizophrenia spectrum disorders (SSD) and healthy controls (HC). By standardizing extraction parameters across the toolkits, we analyzed speech samples from 77 SSD and 87 HC participants and found significant toolkit-dependent variations. While F0 percentiles showed high cross-toolkit correlation ($r=0.962\text{--}0.999$), measures like F0 standard deviation and formant values often had poor, even negative, agreement. Additionally, correlation patterns differed between SSD and HC groups. Classification analysis identified F0 mean, HNR, and MFCC1 (AUC >0.70) as promising discriminators. These findings underscore reproducibility concerns and advocate for standardized protocols, multi-toolkit cross-validation, and transparent reporting.

Index Terms: speech recognition, speech biomarker, clinical speech, acoustic feature extraction

1. Introduction

Acoustic features play a central role in speech processing pipelines, underpinning tasks such as speaker recognition [1, 2], emotion classification [3, 4], and clinical speech assessment [5, 6]. Although end-to-end neural architectures have been increasingly adopted, classical feature-based approaches are still widely used. In clinical contexts, for example, speech signals are often collected in limited quantities due to time, cost, and logistical constraints. Under these conditions, hand-crafted features – whose meaning can be grounded in linguistic and phonetic theory – still provide an interpretable and data-efficient alternative and are sometimes used in combination with more advanced architectures [7, 8].

Despite the continued usage of hand-crafted acoustic features, there is a growing concern that practitioners often use multiple feature sets without fully understanding or validating each feature’s definition, extraction parameters, or applicability. Moreover, the availability of numerous open-source toolkits further complicates the reproducibility and consistency of extracted features. Each toolkit has different underlying assumptions, default settings, and target domains, potentially leading to contradictory or misleading results if not carefully reconciled. In particular, inconsistent or misleading results can have a significant impact in the clinical field, as future studies rely on previous findings to develop screening or monitoring tools for patients.

In this paper, we conducted a comparative analysis of three popular acoustic feature extraction toolkits – OpenSMILE [9],

Praat [10] (via its Python wrapper Parselmouth [11]), and Librosa [12] – applied to a clinical speech dataset of people with Schizophrenia Spectrum Disorders (SSD) and healthy controls (HC). We aligned the parameters as closely as possible across all three tools. This paper aims to answer (1) if acoustic features extracted with different toolkits show consistent results, (2) which acoustic features are most robust across toolkits and participant groups, and (3) what toolkits are reliable to use in clinical studies.

2. Previous Studies

The comparison and validation of acoustic feature extraction tools have been ongoing concerns in speech processing research [13, 14], yet systematic evaluations specifically targeting clinical applications remain limited. Early comparative studies across different toolkits focused primarily on general speech analysis [15, 16], leaving a significant gap in understanding how these tools perform in clinical contexts.

The challenge of reproducibility in clinical speech research has become increasingly apparent, with several studies reporting that different research groups using different tools often produce conflicting results even when analyzing similar populations [17, 18]. This lack of consistency has raised concerns about the reliability of acoustic features as clinical markers and highlighted the need for standardization in feature extraction methodologies.

Recent work has also begun to explore how deep learning approaches might complement or replace traditional acoustic feature extraction [19, 20]. However, the interpretability and theoretical grounding of classical acoustic features continue to make them valuable in clinical contexts, particularly when working with limited data or when explanatory insight is required. This ongoing relevance of traditional acoustic features makes the investigation of their reliability even more critical.

Despite various investigations, there remains a significant gap in our understanding of how different feature extraction tools perform in clinical speech analysis. The present study addresses this gap by providing a systematic comparison of three widely-used toolkits, with a specific focus on features relevant to clinical assessment and the particular challenges posed by pathological speech.

3. Methods

3.1. Data collection

Participants diagnosed with schizophrenia spectrum disorders (SSD; N = 77, females = 24.7%, mean age = 35.92) were enrolled from both inpatient and outpatient departments at a hospital in the US. Participants underwent screening using the psy-

* These authors contributed equally as senior authors.

chosis and mood modules of the Structured Clinical Interview for DSM-IV [21] and were confirmed to meet DSM-5 diagnostic criteria for schizophrenia spectrum disorders. Healthy volunteers ($N = 87$, females = 51.7%, mean age = 36.12) were also enrolled either through their previous involvement in other research studies or by responding to online advertisements. All participants gave informed consent, with minors providing assent. All study procedures were approved by the Institutional Review Board. All participants performed several speech-based tasks, including three picture description tasks, as part of a larger study. All speech samples were digitally recorded, and later manually time-stamped and annotated by trained human annotators.

3.2. Extraction tools

We compared three widely-used acoustic feature extraction toolkits: OpenSMILE, Praat, and Librosa. Each has a distinct implementation approach:

OpenSMILE [9] is designed for batch extraction of large feature sets for machine learning applications. OpenSMILE has been extensively used in various tasks such as speech emotion recognition [22] and clinical speech analysis [17, 23], as well as paralinguistic challenges [24]. We used the eGeMAPS configuration [25] for the extraction, which has become a standard in affective computing and clinical speech research [26].

Praat [10] is primarily designed for interactive phonetic analysis. It employs algorithm-specific implementations for each feature type, with a focus on accuracy over computational efficiency. Praat remains the gold standard in clinical phonetics [17, 27] and detailed acoustic-phonetic studies in the broader Linguistics field [28], particularly when precise voice quality measurements are needed.

Librosa [12] is a Python package originally designed for music information retrieval but increasingly used for speech processing. It employs probabilistic approaches for most of its features and emphasizes perceptual relevance, with its default settings often differing from speech-specific tools. Recently, Librosa has gained popularity in speech analysis for deep learning applications [29] and has also been used in clinical speech assessment [30, 31] due to its integration with Python-based machine learning frameworks.

The three toolkits were clearly built and optimized for different purposes – Praat for phonetic analysis, OpenSMILE for machine learning applications, and Librosa for music information retrieval. However, all of them have been frequently used in the literature, potentially contributing to mixed results.

3.3. Feature Extraction Configuration

To ensure fair comparison, we standardized the extraction parameters across all toolkits:

- Sampling rate: 16kHz across all toolkits
- Frame size: 60ms (equivalent to 960 samples at 16kHz)
- Hop size: 10ms (160 samples)
- Window function: Hamming window
- Pre-emphasis: Disabled for consistency
- Frequency range: 0-8000Hz (Nyquist frequency)
- F0 search range: 55-1000Hz
- Silence thresholds: -60dB for voice activity detection

All other feature-specific thresholds were matched where possible. Despite our efforts to employ consistent parameters, certain toolkit-specific differences remained unavoidable, such

as different underlying algorithms for F0 extraction (cross-correlation-based in OpenSMILE and Praat versus probabilistic in Librosa). All recordings were processed using the standardized pipeline described above, with identical parameter configurations applied consistently across all toolkits. The feature extraction process generated multiple statistics for each acoustic parameter (means, standard deviations, percentiles), which were then used in our correlation analysis.

3.4. Acoustic Features and their Clinical Relevance

Acoustic features provide an objective means to quantify speech patterns that may be altered in clinical populations. In this section, we describe the key features extracted in our study and discussed their relevance to clinical speech assessment, particularly in the context of schizophrenia.

Fundamental frequency (F0) represents the lowest frequency of vocal fold vibration during phonation. F0 characteristics are crucial in clinical linguistics as they reflect both physiological aspects of voice production and prosodic patterns that may be altered in various conditions. In schizophrenia, research has demonstrated abnormal prosodic patterns, often marked by reduced pitch variability and monotonous speech [32, 33].

Formants (F1-F3) are local maxima in the spectrum that result from the acoustic resonance of the human vocal tract. Their values mostly correspond to the position of the tongue and other articulators during vowel production. In clinical populations, alterations in formant patterns may reflect abnormalities in articulation precision or stability. Research has shown that individuals with schizophrenia may exhibit shifts in vowel space areas and less stable formant trajectories, potentially due to motor control deficits or cognitive factors affecting speech planning [34]. The relationship between formants – particularly the ratios and distances between F1, F2, and F3 – provides information about overall vocal tract configuration and motor planning and may serve as biomarkers for certain speech disorders [35].

Harmonics-to-Noise Ratio (HNR) quantifies the relative amount of periodic (harmonic) versus aperiodic (noise) components in the voice signal. This measure is particularly relevant for assessing voice quality, with lower values indicating increased breathiness, roughness, or general dysphonia. In schizophrenia research, reduced vocal quality – possibly related to medication effects, smoking prevalence, or neuromotor factors – has been documented using HNR measures [36].

Jitter and **shimmer** represent cycle-to-cycle variations in frequency and amplitude, respectively. These perturbation measures are highly sensitive to neuromotor control of the vocal folds and have been widely used in clinical voice assessment [37]. In schizophrenia, elevated jitter and shimmer values may indicate reduced precision in laryngeal control [38], potentially related to the broader motor coordination issues associated with the disorder.

Amplitude captures the perceived intensity of speech, reflecting both physiological factors (respiratory support, vocal effort) and communicative intent (emphasis, emotional expression). In clinical populations, abnormal amplitude patterns – whether monotonous speech with minimal intensity variation or inappropriate amplitude modulation – can significantly impact communicative effectiveness. Individuals with schizophrenia often exhibit reduced amplitude variation consistent with negative symptoms like flat affect [33].

Mel-frequency cepstral coefficients (MFCCs) provide a compact representation of the speech spectrum that roughly cor-

responds to the auditory system's response. The first MFCC (MFCC1) primarily reflects overall spectral energy distribution, while higher coefficients (MFCC2-4) capture increasingly fine spectral details related to vocal tract configuration and articulation. In schizophrenia research, altered MFCC patterns have been associated with both production differences (potentially related to articulatory precision or vocal tract tension) and perceptual characteristics that listeners identify as "disorganized" speech [39, 40].

The clinical utility of these acoustic features extends beyond simple group discrimination to monitoring treatment effects, predicting functional outcomes, and potentially serving as objective biomarkers. However, this potential can only be realized if the features can be extracted reliably and consistently across different software implementations—the central focus of our present investigation.

3.5. Statistical analysis

The analysis involved several statistical approaches to evaluate consistency and reliability across the three toolkits. Pearson correlation coefficients were calculated between each toolkit pair for every acoustic feature to assess agreement. Statistical significance of correlations was determined using standard t-tests with p-value thresholds of 0.05, 0.01, and 0.001. To evaluate whether correlations differed significantly between toolkit pairs, Fisher's r-to-z transformation was applied, comparing correlation coefficients with rigorous statistical testing. Additionally, the classification potential of features for distinguishing SSD from HC groups was assessed using ROC curve analysis and calculating Area Under Curve (AUC) values, with features showing AUC values above 0.7 considered to have good discrimination potential.

4. Results

The correlation analysis reveals distinct patterns of agreement between the three feature extraction tools. Figure 1 presents the correlation matrix for both the SSD and HC groups after removing outliers at the 25th and 75th percentiles.

For F0 percentile measurements, we observe remarkably high correlations, particularly between OpenSMILE and Librosa ($r=0.993\text{-}0.999$ for SSD group, $p<0.001$; $r=0.962\text{-}0.993$ for HC group, $p<0.001$). The correlation between OpenSMILE and Praat is similarly strong ($r=0.988\text{-}0.994$ for SSD, $p<0.001$; $r=0.809\text{-}0.981$ for HC, $p<0.001$). However, F0 mean shows more moderate correlation between OpenSMILE and Librosa ($r=0.730$ for SSD, $p<0.001$; $r=0.189$ for HC, $p>0.05$), suggesting algorithm-specific differences in handling of unvoiced frames or edge conditions. Most notably, F0 standard deviation exhibits poor correlation between tools, with negative correlations between OpenSMILE and Librosa ($r=-0.536$ for SSD, $p<0.001$; $r=-0.040$ for HC, $p>0.05$) and between Librosa and Praat ($r=-0.197$ for SSD, $p>0.05$; $r=0.144$ for HC, $p>0.05$). This discrepancy likely stems from fundamental differences in the underlying F0 extraction algorithms and how they handle voice onset/offset transitions.

Formant measurements (F1-F3) show inconsistent extraction across toolkits, suggesting substantial differences in formant estimation algorithms across tools.

Harmonics-to-Noise Ratio (HNR) shows moderate correlation between OpenSMILE and Praat ($r=0.649$ for SSD, $p<0.001$; $r=0.813$ for HC, $p<0.001$) and between Librosa and Praat ($r=0.622$ for SSD, $p<0.001$; $r=0.677$ for HC, $p<0.001$).

However, HNR standard deviation exhibits poor correlation between OpenSMILE and Librosa ($r=-0.534$ for SSD, $p<0.001$; $r=-0.374$ for HC, $p<0.001$) and only moderate correlation between OpenSMILE and Praat ($r=0.084$ for SSD, $p>0.05$; $r=0.289$ for HC, $p<0.05$).

Jitter measurements show reasonable correlation ($r=0.629$ for SSD, $p<0.001$; $r=0.512$ for HC, $p<0.001$) between OpenSMILE and Praat. Similarly, shimmer values exhibit good correlation ($r=0.846$ for SSD, $p<0.001$; $r=0.658$ for HC, $p<0.001$) between OpenSMILE and Praat.

Amplitude mean shows strong correlation between OpenSMILE and Librosa ($r=0.892$ for SSD, $p<0.001$; $r=0.869$ for HC, $p<0.001$) but poorer agreement with Praat ($r=-0.052$ for SSD vs Librosa, $p>0.05$; $r=0.425$ for HC vs Librosa, $p<0.001$). Amplitude percentiles demonstrate more consistent correlation across all three tools, particularly for the SSD group ($r=0.847\text{-}0.948$, $p<0.001$). Interestingly, amplitude standard deviation shows much weaker agreement, especially between OpenSMILE and Praat ($r=0.177$ for SSD, $p>0.05$; $r=0.386$ for HC, $p<0.001$).

The correlation patterns for MFCCs vary considerably by coefficient number. MFCC1 mean shows high correlation between OpenSMILE and Praat ($r=0.981$ for SSD, $p<0.001$; $r=0.989$ for HC, $p<0.001$) but low correlation between OpenSMILE and Librosa ($r=0.100$ for SSD, $p>0.05$; $r=0.406$ for HC, $p<0.001$). Higher-order MFCCs (2-4) show more variable correlation patterns, ranging from strong agreement ($r=0.974$ for MFCC2 OpenSMILE vs Praat in HC, $p<0.001$) to negative correlations ($r=-0.240$ for MFCC2 Librosa vs Praat in SSD, $p<0.05$).

Notably, the correlation patterns differ between the SSD and HC groups for several features, with these differences being statistically significant ($p<0.05$) in features such as F0 mean, HNR mean, and amplitude percentiles. Classification analysis revealed that F0 mean, HNR measures, and MFCC1 features had the highest discrimination potential between SSD and HC groups, with AUC values above 0.75, particularly when extracted using OpenSMILE.

5. Discussion & Conclusion

Our findings reveal significant inconsistencies in feature extraction results, despite careful parameter alignment and standardized processing. These results have important implications for clinical speech analysis and highlight several critical considerations for future research.

The observed discrepancies in feature values raise serious concerns about the reproducibility and reliability of acoustic analyses in clinical applications. While some features demonstrated high cross-toolkit correlation, others showed poor or even negative correlations. Such inconsistencies may explain contradictory or mixed findings across studies using different toolkits. Our results highlight the importance of reporting detailed methodological information, including toolkits and parameters employed, to allow the reproducibility of findings in future research.

Of particular importance is the impact these findings have on interpretability in clinical settings. Unlike consumer applications where end-performance may be the primary concern, clinical contexts demand transparency and interpretability. Medical professionals need to understand why a classification or assessment was made to inform treatment decisions and communicate with patients. When acoustic features are inconsistently extracted, derived conclusions become questionable, regardless of classification accuracy. The interpretability of acoustic features

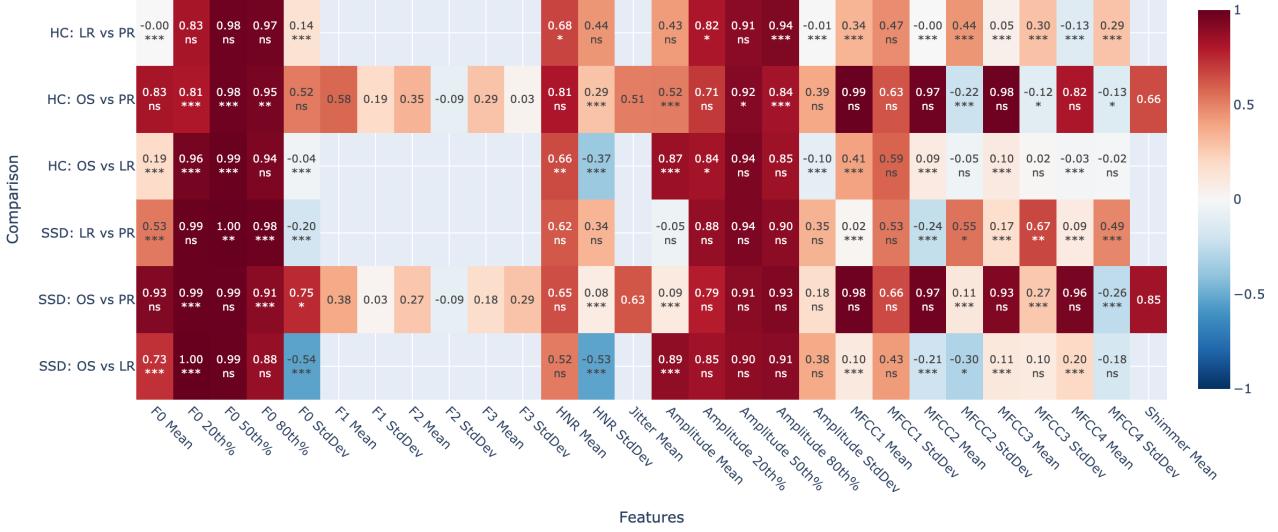


Figure 1: Correlation heatmap comparing acoustic feature extraction across three tools (OpenSMILE [OS], Praat [PR], and Librosa [LR]) for SSD and HC groups. Color intensity indicates correlation strength from -1 (dark blue) to 1 (dark red). Statistical significance of correlations is marked: ns (not significant), * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). Empty cells indicate toolkit pairs not available in specific features.

– their grounding in physiological and linguistic theory – represents one of their key advantages over black-box approaches. However, this advantage is severely compromised when feature extraction itself lacks consistency.

Our results should serve as a call to action for the speech processing community, particularly researchers working in clinical domains. The current practice of extracting hundreds of features and blindly applying dimensionality reduction techniques without proper understanding of the underlying linguistic and physiological basis can be problematic and potentially misleading. This approach may become even more concerning as the field moves toward acoustic embeddings derived from deep neural networks, which rarely offer clinical interpretability. While such techniques may achieve high performance on specific datasets, they may provide little insight into the underlying speech characteristics and may perpetuate or mask extraction inconsistencies.

We recommend that practitioners consider taking the following steps forward:

1. Development of standardized extraction protocols specifically designed for clinical speech analysis, with validated parameters across different toolkits
2. Increased transparency in research publications about extraction methods, tool versions, and parameter configurations to improve reproducibility
3. Cross-validation of acoustic features using multiple extraction tools and datasets before drawing clinical conclusions
4. Greater collaboration between speech technology experts, linguists, and clinical practitioners to ensure feature interpretation is grounded in both technical accuracy and clinical relevance
5. Critical evaluation of newer embedding approaches in clin-

ical contexts, with careful consideration of the trade-off between performance and interpretability

The field stands at a critical juncture where computational advances must be balanced against clinical needs for transparency and theoretical grounding. As automated speech analysis continues to gain traction in healthcare applications, ensuring reliable, interpretable, and consistent measurements becomes imperative. This work contributes to that goal by highlighting current limitations and calling for increased transparency and standardization in acoustic feature extraction for clinical speech analysis.

6. References

- [1] M. Sambur, “Selection of acoustic features for speaker identification,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 2, pp. 176–182, 1975.
- [2] J. Li, X. Fang, F. Chu, T. Gao, Y. Song, and R. L. Dai, “Acoustic feature shuffling network for text-independent speaker verification.” in *Interspeech*, 2022, pp. 4790–4794.
- [3] J. Rong, G. Li, and Y.-P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” *Information processing & management*, vol. 45, no. 3, pp. 315–328, 2009.
- [4] B. T. Atmaja, A. Sasou, and M. Akagi, “Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion,” *Speech Communication*, vol. 140, pp. 11–28, 2022.
- [5] D. Michaelis, M. Fröhlich, and H. W. Strube, “Selection and combination of acoustic features for the description of pathologic voices,” *The Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.
- [6] V. Mittal and R. Sharma, “Machine learning approach for classification of parkinson disease using acoustic features,” *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 233–239, 2021.

- [7] A. N. Omeroglu, H. M. Mohammed, and E. A. Oral, "Multimodal voice pathology detection architecture based on deep and handcrafted feature fusion," *Engineering Science and Technology, an International Journal*, vol. 36, p. 101148, 2022.
- [8] Y. Pan, B. Mirheidari, Z. Tu, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, and H. Christensen, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification," in *Proceedings of Interspeech 2020*. International Speech Communication Association (ISCA), 2020, pp. 4806–4810.
- [9] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.
- [10] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.
- [11] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.
- [12] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.
- [13] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study," *arXiv preprint arXiv:1305.1145*, 2013.
- [14] P. K. Kurzkar, R. R. Deshmukh, V. B. Waghmare, and P. P. Shrishrimal, "A comparative study of feature extraction techniques for speech recognition system," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 3, no. 12, pp. 18 006–18 016, 2014.
- [15] R. Lenain, J. Weston, A. Shivkumar, and E. Fristed, "Surfboard: Audio feature extraction for modern machine learning," *arXiv preprint arXiv:2005.08848*, 2020.
- [16] T. Özseven and M. Düğenci, "Speech acoustic (spac): A novel tool for speech feature extraction and classification," *Applied Acoustics*, vol. 136, pp. 1–8, 2018.
- [17] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. B. Rutkove, K. Kawabata, S. Bhandari, K. Shelton, C. J. Duncan, and V. Berisha, "Repeatability of commonly used speech and language features for clinical applications," *Digital biomarkers*, vol. 4, no. 3, pp. 109–122, 2020.
- [18] S. A. Almaghrabi, D. Thewlis, S. Thwaites, N. C. Rogasch, S. Lau, S. R. Clark, and M. Baumert, "The reproducibility of bio-acoustic features is associated with sample duration, speech task, and gender," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 167–175, 2022.
- [19] P. Best, S. Paris, H. Glotin, and R. Marixer, "Deep audio embeddings for vocalisation clustering," *Plos one*, vol. 18, no. 7, p. e0283396, 2023.
- [20] Q.-B. Hong, C.-H. Wu, H.-M. Wang, and C.-L. Huang, "Combining deep embeddings of acoustic and articulatory features for speaker identification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7589–7593.
- [21] M. B. First and M. Gibbon, "The structured clinical interview for DSM-IV axis I disorders (SCID-I) and the structured clinical interview for DSM-IV axis II disorders (SCID-II)," in *Comprehensive Handbook of Psychological Assessment, Vol. 2. Personality Assessment*, M. J. Hilsenroth and D. L. Segal, Eds. John Wiley & Sons, Inc., 2004, pp. 134–143.
- [22] H. H. Mustafa, N. R. Darwish, and H. A. Hefny, "Automatic speech emotion recognition: a systematic literature review," *International Journal of Speech Technology*, pp. 1–19, 2024.
- [23] H. Lin, C. Karjadi, T. F. Ang, J. Prajarkta, C. McManus, T. W. Alhanai, J. Glass, and R. Au, "Identification of digital voice biomarkers for cognitive health," *Exploration of medicine*, vol. 1, p. 406, 2020.
- [24] A. Ashok, J. Pawlak, S. Paplu, Z. Zafar, and K. Berns, "Paralinguistic cues in speech to adapt robot behavior in human-robot interaction," in *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2022, pp. 01–06.
- [25] 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, 2015.
- [26] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Towards automatic assessment of voice disorders: A clinical approach." in *INTERSPEECH*, 2020, pp. 2537–2541.
- [27] Y. Maryn, "Practical acoustics in clinical voice assessment: a praat primer," *Perspectives of the ASHA Special Interest Groups*, vol. 2, no. 3, pp. 14–32, 2017.
- [28] W. Styler, "Using praat for linguistic research," *University of Colorado at Boulder Phonetics Lab*, 2013.
- [29] M. Bhavya and M. Anal, "Deep learning approach for sound signal processing," in *2022 International Conference on Futuristic Technologies (INCOFT)*. IEEE, 2022, pp. 1–4.
- [30] J. Huang, Y. Zhao, Z. Tian, W. Qu, X. Du, J. Zhang, Y. Tan, Z. Wang, and S. Tan, "Evaluating the clinical utility of speech analysis and machine learning in schizophrenia: A pilot study," *Computers in Biology and Medicine*, vol. 164, p. 107359, 2023.
- [31] R. Banks, C. Higgins, B. R. Greene, A. Jannati, J. Gomes-Osman, S. Tobyne, D. Bates, and A. Pascual-Leone, "Clinical classification of memory and cognitive impairment with multimodal digital biomarkers," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 16, no. 1, p. e12557, 2024.
- [32] A. Parola, A. Simonsen, J. M. Lin, Y. Zhou, H. Wang, S. Ubukata, K. Koelkebeck, V. Bliksted, and R. Fusaroli, "Voice patterns as markers of schizophrenia: building a cumulative generalizable approach via a cross-linguistic and meta-analysis based investigation," *Schizophrenia Bulletin*, vol. 49, no. Supplement_2, pp. S125–S141, 2023.
- [33] M. T. Compton, A. Lunden, S. D. Cleary, L. Pauselli, Y. Alolayan, B. Halpern, B. Broussard, A. Crisafio, L. Capulong, P. M. Balducci *et al.*, "The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech," *Schizophrenia research*, vol. 197, pp. 392–399, 2018.
- [34] A. Hogoboom, M. Rouch, D. Lauerman, L. Pauselli, and M. T. Compton, "Initial evidence of vowel space reduction in a subset of individuals with schizophrenia," *Schizophrenia Research*, vol. 255, pp. 158–164, 2023.
- [35] S. Shellikeri, S. Cho, S. Ash, C. Gonzalez-Recober, C. T. McLellan, L. Elman, C. Quinn, D. A. Amado, M. Baer, D. J. Irwin *et al.*, "Digital markers of motor speech impairments in spontaneous speech of patients with als-ftd spectrum disorders," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 25, no. 3-4, pp. 317–325, 2024.
- [36] Q. Zhao, W.-Q. Wang, H.-Z. Fan, D. Li, Y.-J. Li, Y.-L. Zhao, Z.-X. Tian, Z.-R. Wang, Y.-L. Tan, and S.-P. Tan, "Vocal acoustic features may be objective biomarkers of negative symptoms in schizophrenia: A cross-sectional study," *Schizophrenia Research*, vol. 250, pp. 180–185, 2022.
- [37] J. P. Teixeira and A. Gonçalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Computer Science*, vol. 100, pp. 271–279, 2016.
- [38] P. W. Newman, R. W. Harris, and L. M. Hilton, "Vocal jitter and shimmer in stuttering," *Journal of fluency disorders*, vol. 14, no. 2, pp. 87–95, 1989.
- [39] D. Chakraborty, Z. Yang, Y. Tahir, T. Maszczyk, J. Dauwels, N. Thalmann, J. Zheng *et al.*, "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)*. IEEE, 2018, pp. 6024–6028.

- [40] J. Zhang, P. A. N. Zhongde, G. U. I. Chao, Z. H. U. Jie, and C. U. I. Donghong, “Clinical investigation of speech signal features among patients with schizophrenia,” *Shanghai Archives of Psychiatry*, vol. 28, no. 2, p. 95, 2016.