Speech Characteristics as a Proxy for Depression Severity:

A Building Block for

Future Adaptive Ambulatory Assessment Systems

Zur Erlangung des akademischen Grades einer
DOKTORIN DER PHILOSOPHIE (Dr. phil.)

von der KIT-Fakultät für Geistes- und Sozialwissenschaften des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von

M. Sc. Lisa-Marie Hartnagel (geb. Wadle)

KIT-Dekan: Prof. Dr. Michael Mäs

1. Gutachter: Prof. Dr. Ulrich Ebner-Priemer
2. Gutachter: Prof. Dr. Christof Weinhardt

Tag der mündlichen Prüfung: 18.12.2024

Lisa-Marie Hartnagel (geb. Wadle)

Mental mHealth Lab

Institute of Sports and Sports Science

Karlsruhe Institute of Technology (KIT)

Hertzstr. 15, building 06.31, 76187 Karlsruhe, Germany

lisa.hartnagel@kit.edu

Supervisor: Prof. Dr. Ulrich Ebner-Priemer

SUMMARY

Digital monitoring tools are the most promising approaches to automatically detect impending depressive episodes. Leveraging on the ubiquitous use of smartphones, they open new avenues to continuously collect and objectively extract behavioral markers of depression in the daily life of patients. Human speech characteristics have been considered as a potential behavioral marker for this endeavor, as a growing number of case-control studies show depression-related changes in acoustic and linguistic speech features. However, longitudinal studies are sparse but necessary to understand whether within-person changes in speech characteristics could be used for the early identification of prodromal symptoms, and thus the prevention of new clinical episodes.

In the present doctoral thesis, I investigate the informative value of acoustic and linguistic features extracted from longitudinally collected everyday speech samples with respect to depression severity. Articles 1-3 are based on a dataset of 30 patients with an acute depressive episode undergoing sleep deprivation therapy. Before, during, and after treatment patients collected speech samples and reported current symptom severity 2-3 times per day by ambulatory assessment. Sleep deprivation therapy allows to observe treatment effect and relapse within four days, thus the dataset contains a wide range of depressive state levels.

Using multilevel regression models, I revealed associations between three preselected acoustic features and depression severity. Specifically, lower pitch variability, higher speech rate, and shorter speech pauses were associated with lower depression severity and more pleasant momentary states in general. A combined model of all three speech features explained 2% of variance in depression severity. Article 2 extends the depression-related findings of Article 1 by using multi-parameter machine learning models including 89 speech features and different train-test split scenarios. The superior model tested explained 33.9% of the variance in depression severity and the results further suggest the need for personalized models. In Article 3, I shifted my focus to linguistic analysis and identified more positive emotion words and fewer negative emotion words to be associated with lower reported depression severity. The commentary presented afterwards introduces the term *smart digital phenotyping* and discusses the challenges when extracting a large number of features from behavioral markers and data protection concerns. Concluding, a discussion of the main findings and limitations, followed by an outlook on future research avenues can be found. In summary, this doctoral thesis represents a building block for future speech-based adaptive ambulatory assessment systems that could one day be used to monitor depression.

ZUSAMMENFASSUNG

Digitales Monitoring ist ein vielversprechender Ansatz, um drohende depressive Episoden automatisch zu erkennen. Die allgegenwärtige Nutzung von Smartphones eröffnet dabei neue Möglichkeiten zur kontinuierlichen Erfassung und objektiven Extraktion von depressionsrelevanten Verhaltensmarkern im Alltag der Patienten.

Menschliche Sprachmerkmale werden als potenzielle Verhaltensmarker für dieses Vorhaben in Betracht gezogen, da eine wachsende Zahl von Case-Control-Studien depressionsbezogene Veränderungen in akustischen und linguistischen Sprachmerkmalen zeigen. Es gibt jedoch noch wenige Längsschnittstudien, die notwendig wären, um zu verstehen, ob Veränderungen von Sprachmerkmalen auch innerhalb einer Person zur Früherkennung von Prodromalsymptomen und damit zur Vorbeugung neuer klinischer Episoden genutzt werden können.

In der vorliegenden Dissertation untersuche ich die Aussagekraft von akustischen und linguistischen Merkmalen in Bezug auf den Schweregrad der Depression. Die Sprachmerkmale wurden aus Sprachproben extrahiert, die von den Patienten in ihrem Alltag längsschnittlich gesammelt wurden.

Artikel 1-3 basieren auf einem Datensatz von 30 Patienten mit einer akuten depressiven Episode, die sich einer Schlafentzugstherapie unterzogen. Vor, während und nach der Behandlung sammelten die Patienten 2-3 Mal pro Tag per Ambulantem Assessment Sprachproben und gaben zusätzlich Ratings über ihre aktuelle Symptomschwere ab. Die Schlafentzugstherapie ermöglicht die Beobachtung von Behandlungseffekts und Rückfall innerhalb weniger Tage, so dass der Datensatz ein breites Spektrum an Depressionsleveln enthält.

Mithilfe von Mehrebenenmodellen konnte ich Zusammenhänge zwischen drei vorausgewählten akustischen Merkmalen und dem Schweregrad der Depression aufzeigen. Dabei zeigten sich Zusammenhänge zwischen geringerer Tonhöhenvariabilität, höherer Sprechgeschwindigkeit und kürzerer Sprechpausen mit einer geringeren Depressionsschwere und angenehmeren momentanen Zuständen im Allgemeinen. Ein kombiniertes Modell aus allen drei Sprachmerkmalen ergab eine aufgeklärte Varianz der Depressionsschwere von 2%. Artikel 2 erweitert die depressionsbezogenen Ergebnisse aus Artikel 1 durch die Verwendung von Multiparameter-Modellen für maschinelles Lernen, die 89 Sprachmerkmale und verschieden Aufteilung in Trainings- und Test-Sets enthalten. Das beste Modell erklärte 33.9 % der Varianz des Schweregrads der depressiven Zustände. Weiterhin deuten die Ergebnisse auf die Notwendigkeit personalisierter Modelle hin. In Artikel 3 lag der Fokus auf linguistischen Analysen und ich konnte zeigen, dass ein erhöhter Gebrauch von positiven Emotionswörtern und geringerer Gebrauch von negativen Emotionswörtern mit einem niedrigeren berichteten Depressionsschweregrad verbunden sind. Der anschließende Kommentar führt den Begriff *Smart Digital Phenotyping* ein und erörtert die Herausforderungen bei der Extraktion einer großen Anzahl von Merkmalen aus Verhaltensmarkern sowie Datenschutzbedenken zu finden. Schließlich ist eine Diskussion der wichtigsten Ergebnisse und Einschränkungen, gefolgt von einem Ausblick auf zukünftige Forschungsmöglichkeiten. Zusammenfassend lässt sich sagen, dass diese Doktorarbeit einen Baustein für zukünftige sprachbasierte adaptive ambulante Assessment Systeme darstellt, die eines Tages zum Monitoring von Depressionen eingesetzt werden könnten.

PREFACE

Chapter 2 is based on a peer-reviewed article that has been published as Wadle, L.-M., Ebner-Priemer, U. W., Foo, J. C., Yamamoto, Y., Streit, F., Witt, S. H., Frank, J., Zillich, L., Limberger, M. F., Ablimit, A., Schultz, T., Gilles, M., Rietschel, M., & Sirignano, L. (2024). Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study. *JMIR Mental Health*, 11, e49222. https://doi.org/10.2196/49222

Chapter 3 is based on a manuscript that has been submitted to *JMIR Mental Health* as Hartnagel, L.-M., Emden, D., Foo, J. C., Streit, F., Witt, S. H., Frank, J., Limberger, M. F., Schmitz, S., Gilles, M., Rietschel, M., Hahn, T., Ebner-Priemer, U. W., & Sirignano, L. [under review]. Speech-based Machine Learning for Momentary Depression-Severity Prediction in Acutely Depressed Patients undergoing Sleep Deprivation Therapy

Chapter 4 is based on a peer-reviewed article that has been published as Hartnagel, L.-M., Ebner-Priemer, U. W., Foo, J. C., Streit, F., Witt, S. H., Frank, J., Limberger, M. F., Horn, A. B., Gilles, M., Rietschel, M., & Sirignano, L. (2024). Linguistic style as a digital marker for depression severity: An ambulatory assessment pilot study in patients with depressive disorder undergoing sleep deprivation therapy. *Acta Psychiatrica Scandinavica*, 1-10. https://doi.org/10.1111/acps.13726

Chapter 5 is based on a peer-reviewed commentary that has been published as Wadle, L.-M., & Ebner-Priemer, U. W. (2023). Smart digital phenotyping. *European Neuropsychopharmacology*, 76, 1-2

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to a number of people who have guided and supported me throughout my Ph.D. journey. First and foremost, I would like to thank my supervisor Prof. Dr. Ulrich Ebner-Priemer for giving me the opportunity to complete my doctoral studies and this dissertation under his supervision. Uli, thank you for letting me bring the new world of audio into your lab, and for your patient support and constructive feedback at any time. I appreciate you gave me the freedom and the guidance I needed.

Moreover, I am thankful to Prof. Dr. Christof Weinhardt not only for being my second reviewer, but especially for bringing the KD²School to life. I feel honored to be part of such a stimulating interdisciplinary research training group that has created a wonderful and motivating research atmosphere. I am also grateful to Prof. Dr. Tanja Schultz for supporting me as a second supervisor, in particular with her expertise in speech analysis.

A big thank you to Lea Sirignano and Prof. Dr. Marcella Rietschel for providing me with the SLEDGE dataset. Thank you for being open to collaboration and Lea, for being my patient data cleaning buddy. I am also grateful to all my co-authors for their valuable contributions!

Thank you to all my colleagues in the mHealth Lab, especially Matthias Limberger, for our scientific (often statistical) discussions, recharging lunch breaks, and the many shared laughs. I would also like to thank everyone involved in the KD²School, especially the other doctoral students. Knowing that you are not alone on this roller coaster ride has given me courage and strength. My special thanks goes to Hannah Seidler, who has become a true friend. Thank you for your open ears and warm heart!

My wonderful family and friends deserve a special mention. Your support and encouragement have carried me through some dry spells and have made celebrations even more enjoyable. To my parents: I am forever grateful to you for equipping me with curiosity and always believing in me. And to Moritz, thank you for being by my side, solid as a rock.

TABLE OF CONTENTS

List of Abbreviations

AA	ambulatory assessment
AI	artificial intelligence
ADS-K	Allgemeine Depressionsskala Kurzfom
DSM	Diagnostic and Statistical Manual of Mental Disorders
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set
EAR	Electronically Activated Recorder
e-diaries	electronic diaries
EU	European Union
F0	fundamental frequency
ICC	intraclass correlation coefficient
ICD-10	International Classification of Diseases, Tenth Revision
JITAI	just-in-time adaptive intervention
LIWC	Linguistic Inquiry and Word Count
LOSO	leave-one-subject-out
MADRS	Montgomery–Åsberg Depression Rating Scale
MAE	mean absolute error
MDD	major depressive disorder
MDMQ	Multidimensional Mood Questionnaire

MFCC	mel-frequency cepstral coefficients
ML	machine learning
openSMILE	open-source Speech and Music Interpretation by Large-space Extraction
RQ	research question
SDT	sleep deprivation therapy
SLEDGE	Sleep Deprivation and Gene Expression Project
SVR	support vector regression
XGBoost	eXtreme Gradient Boosting regression

LIST OF FIGURES

LIST OF TABLES

O, how wonderful is the human voice!

It is indeed the organ of the soul!

from *Hyperion: a romance*
by Henry W. Longfellow

CHAPTER 1

GENERAL INTRODUCTION

Major Depressive Disorder

Major depressive disorder (MDD) is a severe mental disorder that is the third leading cause of disease burden worldwide and is expected to rank first by 2030 (Malhi & Mann, 2018). The global lifetime prevalence of MDD is extremely high, estimated at 30-40% based on a prospective epidemiologic study (Moffitt et al., 2010). Nearly 230 million people worldwide were diagnosed with MDD in 2021, representing approximately 3% of the world's population (Global Burden of Disease Collaborative Network, 2020). An estimated increase of an additional 53 million cases was reported due to the COVID-19 pandemic (Santomauro et al., 2021). The societal economic burden caused by MDD was valued to be $333.7 billion US dollars in 2019 (Greenberg et al., 2023).

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; American Psychiatric Association, 2013), MDD is characterized by a range of different symptoms, with low mood and a loss of interest in daily activities at its core being present for the same 2-week time period. A list of symptoms can manifest negatively, including but not limited to issues with sleep quality, appetite, self-worth, guilt, concentration, and physical activity, also and suicidal ideation can occur (American Psychiatric Association, 2013). Specifically, according to widely used diagnostic tools, the experience of 400 different symptom combinations can be labeled as MDD (Goldberg, 2011; Østergaard et al., 2011), forming heterogeneous clinical pictures (Fried, 2017; Fried & Robinaugh, 2020). Moreover, MDD is associated with negative outcomes, such as job loss (Lerner et al., 2004; Vos et al., 2020), premature mortality (Miloyan & Fried, 2017), cardiovascular disease (Hare et al., 2014), and reduced quality of life (Saragoussi et al., 2018).

Significant scientific breakthroughs that also show in decreased prevalence or burden of disease statistics have not been announced over the past three decades (Jorm et al., 2017). This may be due to the complex interplay of contributing factors, and the lack of clear biomarkers that could be used for objective testing or treatment monitoring, with the consequence of having to rely on subjective self-reports (Kapur et al., 2012; Rimti et al., 2023; Scull, 2021; Thibaut, 2018). The etiology of MDD is still object of research, but it is assumed to be multifactorial, involving biological, environmental, psychosocial, and genetic factors (Malhi & Mann, 2018).

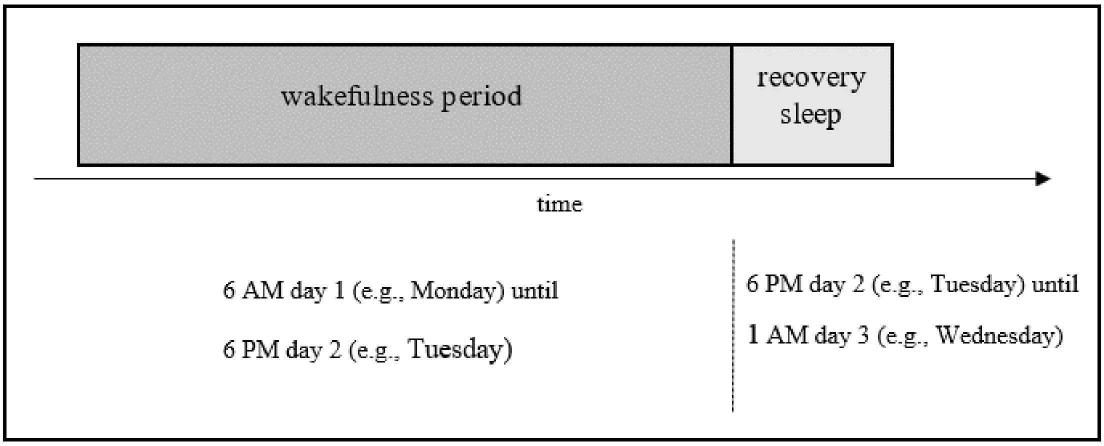
While some patients experience a single episode in their lifetime, MDD evolves as a chronic and relapsing illness with fluctuating levels of depression severity in many cases (Verduijn et al., 2017). Approximately 50% of patients with MDD experience an ebb and flow of relapsing and remitting periods during their lifetime (Verduijn et al., 2017). Therefore, secondary prevention, i.e., relapse prevention is an important treatment goal (Benasi et al., 2021). This requires long-term care and monitoring to identify early indicators of relapse and to track treatment response. There are several treatment options available for MDD (Malhi & Mann, 2018), the discussion of which is beyond the scope of this doctoral thesis. Therefore, I focus on detailing sleep deprivation therapy (SDT) as the treatment implemented in the dataset used in my doctoral thesis.

SDT is a chronotherapeutic intervention, in which patients are temporarily deprived of sleep (Boland et al., 2017; Dallaspezia & Benedetti, 2015; Demet et al., 1999;

Wirz-Justice & Benedetti, 2020). Specifically, patients are kept awake for about 36 hours, as shown in Figure 1 (Dallaspezia & Benedetti, 2015). Approximately 60% of patients experience a reduction in depressive symptoms within hours in the morning following the night of sleep deprivation (Dallaspezia & Benedetti, 2015; Wirz-Justice & Van Den Hoofdakker, 1999). This rapid onset of treatment is a major advantage over antidepressants, which can take up to weeks to take effect. Another advantage of this therapy approach is that it has virtually no side effects, except for daytime sleepiness.

Figure 1

Exemplary Sleep Deprivation Therapy


wakefulness period
recovery sleep
time
6 AM day 1 (e.g., Monday) until
6 PM day 2 (e.g., Tuesday)
6 PM day 2 (e.g., Tuesday) until
1 AM day 3 (e.g., Wednesday)

Unfortunately, the treatment effect is transient with about 80% of patients relapsing after a night of recovery sleep (Dallaspezia & Benedetti, 2015; Leibenluft & Wehr, 1992). Short-term relapse can be attenuated with strategies such as light therapy or sleep phase advance (Echizenya et al., 2013), but only for a minority of patients, achieved effects remain (Giedke & Schwärzler, 2002). The complex underlying mechanism of SDT is not yet understood, but probably involves an altered brain metabolism caused by the lack of sleep (Dallaspezia & Benedetti, 2015).

In general, prerequisite for adequate prediction and treatment of MDD is the ability to measure it (Fried et al., 2022). The current state-of-the-art MDD assessment relies on self-report in the form of interviews, questionnaires, and rating scales. A collection of more than 280 measures has been described in the scientific literature (Santor et al., 2006). Individuals are asked to rate how severely they have been affected by a list of depression symptoms over a predefined period of time. Equivalent to the DSM-5 criteria stating that symptoms have to be present for a time frame of two weeks (American Psychiatric Association, 2013), this duration is also often referred to in the assessments (Colombo et al., 2019). Despite being the current gold standard, these self-report tools do not come without drawbacks.

First, self-report measures are subjective and rely on retrospective recall. Patients have to recall emotional, behavioral, or cognitive symptoms, a process prone to systematic bias (e.g., Ebner-Priemer & Trull, 2009; Stone et al., 2007). For example, the mood congruency effect describes an easier retrieval of information that is consistent with the current emotional state, and the peak-end rule states that recall is often dominated by the most intense and the most recent experiences (Kahneman et al., 1993; Kihlstrom et al., 2000). Thus, retrospective reporting in MDD may be blurred by cognitive biases and dysfunctional perceptions rather than accurately reflecting current symptom severity or mirroring actual experiences over, say, the previous two weeks (A. G. Horwitz et al., 2023; Wells & Horwood, 2004; Zupan et al., 2017). Second, assessments are infrequent. Mostly administered during clinical visits at arbitrary points in time, assessments do not reflect the natural course of symptoms, thus crucial information may be lost (Ebner-Priemer & Santangelo, 2020; Ebrahimi et al., 2021). Patients are often required to summarize their symptoms and experiences into a single response or score, which fails to capture the dynamic ebb and flow of symptoms and fluctuations in affective states (Ebner-Priemer & Santangelo, 2020; Ebrahimi et al., 2021). This is also reflected in MDD studies, showing limited congruence between retrospective symptom reports and actual dynamics (Solhan et al., 2009; Wells & Horwood, 2004). Even the detection of treatment response may be delayed when relying solely on sporadic assessments. By comparing weekly assessment interviews and daily process methods, Lenderking and colleagues (2008) demonstrated that daily methods can detect treatment effects more quickly than standard assessments, where the earliest time to detect therapeutic impact is the next clinical visit.

Ambulatory Assessment in Major Depressive Disorder

To overcome these limitations, methodological developments leveraging on technological advances are promising. Ambulatory assessment (AA; Fahrenberg, 1996; Fahrenberg et al., 2007), also known as experience sampling method (Larson & Csikszentmihalyi, 1983), ecological momentary assessment (Stone & Shiffman, 1994), and digital phenotyping (Torous et al., 2016), has become the gold standard for studying individuals in their daily lives. Although the terms differ, the methodologies are united by the use of personal digital devices, such as smartphones, electronic diaries (e-diaries), and wearables, to repeatedly assess human behavior, symptoms, and physiological and biological processes as they unfold in daily life as individuals go about their normal daily activities (Trull & Ebner-Priemer, 2013).

The main advantages of AA are the ability to collect real-life data in real time, minimizing retrospective recall bias and bridging the gap between clinical visits while increasing ecological validity (e.g., Ebner-Priemer & Trull, 2009; Trull & Ebner-Priemer, 2014). In addition, AA allows to capture dynamic within-person changes because data are collected repeatedly; the patient does not have to average experiences across time and situations (e.g., Ebner-Priemer & Trull, 2009; Trull & Ebner-Priemer, 2013). Furthermore, various types of data can be collected objectively, continuously, passively and thus unobtrusively over a prolonged period of time using smartphones or additional wearables (e.g., Reichert et al., 2021; Torous et al., 2017). For example, social withdrawal could be estimated using GPS information, and instead of asking people about their sleep quality over the past two weeks, a sleep tracker could be informative and provide objective data. Additionally, multimodal data collection is possible, and combining active (e.g., questionnaire via e-diary) and passive (e.g., physical activity) data streams can even provide a more holistic picture of the daily life of an individual (e.g., Matcham et al., 2022). Moreover, AA allows to capture contextual specificities of an individual's natural environment, such as the identification of specific stressors (Trull & Ebner-Priemer, 2013). Based on this information, so-called just-in-time adaptive interventions (JITAIS) and interactive feedback are feasible, allowing the right intervention to be offered at the right time (Nahum-Shani et al., 2018).

As a result, avenues are opened for a shift from reacting with treatment after diagnosis to preventing a full-blown episode; in other words, intervening when symptoms are still sub-threshold. This could greatly improve patient management and care. Importantly, these new approaches do not replace traditional ones, but rather serve as a diagnostic adjuncts. Given the ubiquitous use of smartphones today, AA is a promising tool for the timely detection of behavioral, cognitive, and physiological changes that occur in the trajectory of mental health disorders (Ebner-Priemer & Santangelo, 2020; Onnela & Rauch, 2016; Trull & Ebner-Priemer, 2014). Trull and Ebner-Priemer (2013) discuss the potential for the mental health context, which includes a) investigating symptom dynamics and mechanisms, b) predicting future symptom reoccurrence or onset, c) tracking of treatment effects, and d) predicting treatment effects. With regard to MDD, there is a body of AA research that discusses various indicators of daily life associated with depressive symptoms, such as physical activity, location, phone use, and speech data (for reviews see De Angel et al., 2022; Zarate et al., 2022).

Human Speech as a Proxy for MDD Severity

Recently, a person’s speech has been increasingly studied in relation to MDD. When we speak, we express our thoughts, intentions, and emotions, providing a wealth of information and a window into our innermost selves (Jablonka et al., 2012; Kappas et al., 1991; Low et al., 2020; Pennebaker et al., 2003; Tausczik & Pennebaker, 2010). Speech characteristics can be broadly divided into two facets: acoustic and linguistic features, put simply, the *how* a person speaks and the *what* a person says. Acoustic features include all features that reflect mathematical properties of the sound wave, describing prosody (e.g., intonation), tempo (e.g., pause time), loudness (e.g., volume), and voice quality (e.g., jitter), among others (Koops et al., 2023). The most prominent computational tool to objectively extract acoustic features is the software *open-source Speech and Music Interpretation by Large-space Extraction* (openSMILE; Eyben et al., 2010). Linguistic features refer to the content of speech or specific word usage (e.g., thematic word categories, grammar, tenses, vocabulary diversity) and require a transcript of speech samples. The dictionary-based Linguistic Inquiry and Word Count (LIWC) is widely used for linguistic analysis (Pennebaker et al., 2015).

The idea that speech-based information could serve as a proxy for depressive states dates back to early clinical observations by Emil Kraepelin in 1921. He described the voice of MDD patients as low, hesitant, slow, monosyllabic, and monotonous (Kraepelin, 1921). After decades of research, it is now assumed that MDD symptoms can have a number of different effects on the highly complex neuromuscular speech production network, which involves over 100 muscles (Cannizzaro et al., 2004; Low et al., 2020). Specifically, imaging methods and clinical observations suggest the involvement of the motor cortex, the supplementary motor area, the basal ganglia, and the cerebellum (Wildgruber et al., 2001). Dopaminergic changes typical of MDD are thought to directly affect basal ganglia structures responsible for the motor control of speech movements (Wildgruber et al., 2001). Furthermore, typical MDD symptoms such as reduced cognitive functioning, fatigue, and persistent negative affect may impact speech planning and production (Caligiuri & Ellwanger, 2000). These alterations can then be estimated by various parameters of the acoustic waveform (Eyben et al., 2016). In terms of linguistic style, presumed MDD-related changes have been motivated by theories of depression. For example, specific characteristics in word use have been derived from Beck’s Cognitive Model of Depression (Beck et al., 1979) or heightened self-focus theories (Pyszczynski & Greenberg, 1987). Deficits in cognitive and executive functioning are also discussed as possible reasons (Trifu et al., 2017).

Recent reviews provide an overview of the potential of speech characteristics in relation to a variety of psychiatric disorders (Low et al., 2020), and depression in particular (Cummins et al., 2015; Koops et al., 2023). However, much of this evidence is based on case-control designs (Low et al., 2020). These between-person comparisons contribute to the understanding of speech-related alterations within a person only to a limited extend. However, our ultimate goal is to monitor a patient’s fluctuating symptoms over time to identify prodromal signs, e.g., of an impending episode. Therefore, we need patient data collected at a more granular level, and collected longitudinally, ideally before an episode, during an emerging episode showing prodromal symptoms, at best also during, and after a fully developed episode. This would open new avenues for monitoring MDD symptoms, not only allowing to detect episodes once they have occurred, but more importantly, to identify prodromal patterns, with the chance to prevent relapse (Ebrahimi et al., 2021; Fried et al., 2022).

Particularly in an AA context, everyday speech data as a proxy for depression severity is a promising and advantageous candidate for many reasons: Most of us speak naturally in our everyday lives, and most of us also carry a smartphone with a built-in microphone with us almost all the time. Given informed consent, this opens up the possibility of collecting speech remotely and in a cheap, simple, noninvasive, and unobtrusive manner (Koops et al., 2023). In addition, data can be collected continuously and advances in computational analysis allow for objective extraction and analysis of speech features.

However, there is only a small number of longitudinal studies on speech in MDD conducted in a clinical population that provides first, though not always consistent, insights relevant to this endeavor, beyond findings from between-person studies (Arevian et al., 2020; Campbell et al., 2023; Cummins et al., 2023; Gerczuk et al., 2022; R. Horwitz et al., 2013; Mundt et al., 2007, 2012; Quatieri & Malyska, 2012; Stiles et al., 2023; Trevino et al., 2011; Yang et al., 2013). Note that work by Horwitz et al. (2013), Quatieri and Malyska (2012), and Trevino et al. (2011) are all based on the same dataset by Mundt and colleagues (2007), and Campbell et al. (2023) and Cummins et al. (2023) also use the same underlying dataset. Depending on the research questions, I will present the relevant studies, their results, strengths, and limitations in the respective articles in Chapter 2-4. To sum up, a first limiting factor of most of these studies is the rather long time intervals between assessments, hindering to capture dynamic fluctuations. In detail, speech analysis is based on a speech sample collected every week (Arevian et al., 2020; R. Horwitz et al., 2013; Mundt et al., 2007; Quatieri & Malyska, 2012), every other week (Campbell et al., 2023; Cummins et al., 2023; Trevino et al., 2011), every seven weeks (Yang et al., 2013), or is even limited to a pre-post treatment comparison (Mundt et al., 2012), even when collected on a higher frequency (Stiles et al., 2023). Overall, only Gerczuk and colleagues (2022) aimed to collect data on a more dynamic level, asking for three speech samples per day over two weeks. Second, the analysis methods are mixed (e.g., correlations, machine learning) but rarely do justice to the data structure at hand, which is repeated and thus nested data, requiring multilevel regression models. As a result, the strengths of the longitudinal data structure are not fully exploited.

Research Questions

In an effort to unlock the full potential of everyday speech data, I aimed to contribute a building block on the journey towards a speech-based depression monitoring system in this doctoral thesis. My goal was to explore human speech characteristics extracted from longitudinally collected speech samples and their associations to momentary depressive states in daily life. In Article 1 and 2, I analyze the link between acoustic features and momentary depressive states. In Article 3, I investigate linguistic features in association to momentary depressive states. In form of a commentary, Article 4 concludes with a broader discussion on challenges and chances of digital phenotyping.

Articles 1-3 are based on the same dataset. In short, a sample of 30 inpatients with an acute depressive episode underwent SDT. Before, during, and after therapy, patients were asked via AA to report how they currently feel 2-3 times per day and to record a selfie video at concomitant time points. The audio tracks of these selfie videos served as speech samples from which acoustic and linguistic features were extracted with state-of the art computational tools. A major benefit of SDT is that treatment effect and relapse can be observed within a few days, which allowed to investigate acoustic and linguistic features in the context of varying momentary depressive states.

In Article 1, I analyze the associations between a pre-defined set of three acoustic speech features and momentary depressive states. Based on previous research findings, I focused on the features pitch variability, speech rate, and speech pauses to avoid alpha-error inflation. Besides momentary depressive states, I also included more broadly defined momentary affective states in the analyzes, namely negative affect, positive affect, valence, energetic arousal, and calmness. This opened the opportunity to gain insights into the specificity or transdiagnostic nature of associations. Conducting multilevel analysis, I aimed to replicate previous findings regarding the association between speech features and depression severity, and to extent them to within-person data. An additional objective was to evaluate generalizability of results to more broadly defined momentary affective states non-specific to depression.

RQ 1: Do acoustic features reflect within-person variability in momentary depressive states?

In Article 2, I aimed to understand whether a speech-based multi-parameter machine learning approach would improve the depression severity prediction compared to my analyzes in Article 1. In total, a comprehensive set of over 500 machine learning pipelines were run evaluating random forest, linear regression, support vector regression, and eXtreme gradient boosting regression models. A further target of this work was to test five different train-test split scenarios and their impact on prediction performance. Specifically, a group 5-fold cross-validation on subject level, a leave-one-subject-out approach, a chronological split, an odd-even split, and a random split were tested.

RQ 2: Does speech-based multi-parameter machine learning enhance predictive performance for depression severity, and which role do train-test splitting techniques play?

In Article 3, I switch my focus to linguistic analysis. Using LIWC, I extracted word categories from transcribed speech samples and explored the usage of positive and negative emotions words, first-person pronouns, and past tense words relative to depression level.

RQ 3: Does linguistic style reflect within-person variability in momentary depressive states?

Additionally, in a commentary, I discuss that digital phenotyping comes with the challenge of generating and handling a plethora of features that often lack relevance to the clinical phenomena being studied. Introducing the term *smart digital phenotyping*, I advocate for features closely resembling psychopathology rather than being easy-to-access, and to extract them in a smart, privacy preserving way.

In the present Chapter 1, I provided background information on MDD, AA, and human speech characteristics, and conclude with three RQs that will be addressed in the following chapters. Chapters 2-5 comprise four peer-reviewed articles, which are all but one published. Article 2 in Chapter 3 is under review. Chapter 6 comprehensively summarizes the main results of this doctoral thesis, followed by a discussion of limiting aspects, and an outlook on future research avenues. The reference list and the appendix build the end of this doctoral thesis.

CHAPTER 2

ARTICLE 1:

SPEECH FEATURES AS PREDICTORS OF MOMENTARY DEPRESSION SEVERITY

This chapter is based on an adapted version of the peer-reviewed article published as

Wadle, L.-M., Ebner-Priemer, U. W., Foo, J. C., Yamamoto, Y., Streit, F., Witt, S. H., Frank, J., Zillich, L., Limberger, M. F., Ablimit, A., Schultz, T., Gilles, M., Rietschel, M., & Sirignano, L. (2024). Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study. *JMIR Mental Health*, 11, e49222. https://doi.org/10.2196/49222

Abstract

Background: The use of mobile devices to continuously monitor objectively extracted parameters of depressive symptomatology is seen as an important step in the understanding and prevention of upcoming depressive episodes. Speech features such as pitch variability, speech pauses, and speech rate are promising indicators, but empirical evidence is limited, given the variability of study designs.

Objective: Previous research studies have found different speech patterns when comparing single speech recordings between patients and healthy controls, but only a few studies have used repeated assessments to compare depressive and non-depressive episodes within the same patient. To our knowledge, no study has used a series of measurements within patients with depression (e.g., intensive longitudinal data) to model the dynamic ebb and flow of subjectively reported depression and concomitant speech samples. However, such data are indispensable for detecting and ultimately preventing upcoming episodes.

Methods: In this study, we captured voice samples and momentary affect ratings over the course of three weeks in a sample of patients (*N*=30) with an acute depressive episode receiving stationary care. Patients underwent sleep deprivation therapy, a chronotherapeutic intervention that can rapidly improve depression symptomatology. We hypothesized that within-person variability in depressive and affective momentary states would be reflected in the following 3 speech features: pitch variability, speech pauses, and speech rate. We parametrized them using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) from open-source Speech and Music Interpretation by Large-Space Extraction (openSMILE) and extracted them from a transcript. We analyzed the speech features along with self-reported momentary affect ratings, using multilevel linear regression analysis. We analyzed an average of 32 (*SD* 19.83) assessments per patient.

Results: Analyses revealed that pitch variability, speech pauses, and speech rate were associated with depression severity, positive affect, valence, and energetic arousal; furthermore, speech pauses and speech rate were associated with negative affect, and speech pauses were additionally associated with calmness. Specifically, pitch variability was negatively associated with improved momentary states (i.e., lower pitch variability was linked to lower depression severity as well as higher positive affect, valence, and energetic arousal). Speech pauses were negatively associated with improved momentary states, whereas speech rate was positively associated with improved momentary states.

Conclusion: Pitch variability, speech pauses, and speech rate are promising features for the development of clinical prediction technologies to improve patient care as well as timely diagnosis and monitoring of treatment response. Our research is a step forward on the path to developing an automated depression monitoring system, facilitating individually tailored treatments and increased patient empowerment.

Introduction

Background

Depression is one of the most prevalent health disorders worldwide (Streit et al., 2023; Vos et al., 2020). The World Health Organization predicted that depression would be one of the three leading causes of disease burden by 2030 (Mathers & Loncar, 2006), even before its prevalence increased owing to the COVID-19 pandemic (Santomauro et al., 2021). This disorder has symptoms that include depressed mood, loss of energy and interest, sleep problems, and diminished ability to concentrate (American Psychiatric Association, 2013); thus, depression imposes a substantial burden on the patients as well as their surroundings, society, and the economy (Sobocki et al., 2006). Most importantly, depression is a chronic disorder, characterized by multiple episodes over years or decades. However, strategies for secondary prevention or early detection of new episodes are missing.

The diagnosis and severity assessment of depression relies mostly on self- or caregiver reports, which are prone to retrospective and social desirability bias (Ben-Zeev & Young, 2010; Eaton et al., 2000). In addition, such assessments are time and resource intensive because clinical specialists are needed over the course of treatment and recovery (Abd-Alrazaq et al., 2023). Moreover, many new episodes remain undiagnosed or untreated, that is, secondary prevention is the main issue (Kohn et al., 2004; Williams et al., 2017). To reduce burden, the timely detection and diagnosis of (new) depressive episodes are critical.

In recent years, research has focused on the identification of mental health disorder indicators that can be derived automatically, driven by technological developments (De Angel et al., 2022; Torous et al., 2015). In particular, the innovation of the ambulatory assessment research technique has contributed strongly to this endeavor (Zarate et al., 2022). Different terms have been used for this kind of methodology: ambulatory assessment (Fahrenberg et al., 2007), ecological momentary assessment (Stone & Shiffman, 1994), experience sampling (Csikszentmihalyi & Larson, 1987), and digital phenotyping (Torous et al., 2016). Although the terms differ, all approaches use computer-assisted methodology to assess momentary self-reported symptoms (e.g., via electronic diaries (e-diaries)), behaviors, or physiological processes, or actively or passively collect smartphone and physical data or context information (e.g., via wearables) while the participant performs normal daily activities in their natural environment (Ebner-Priemer

& Trull, 2009). The main advantages of ambulatory assessment are (1) the ability to collect real-life data in real time, thereby reducing retrospective recall bias and increasing ecological validity; and (2) the ability to collect data continuously (passively), which allows us to capture dynamic changes. Accordingly, ambulatory assessment is a promising tool for the timely detection of upcoming clinical episodes to prevent further clinical deterioration (Ebner-Priemer & Santangelo, 2020; Onnela & Rauch, 2016; Trull & Ebner-Priemer, 2014). In particular, parameters captured objectively by wearables are useful because they can be assessed passively with a high frequency over prolonged time periods (Torous et al., 2017).

Promising markers that can be assessed objectively are speech and language, which are also metaphorically called *the mirror of the soul* (Sundberg, 1998). Even before objective measurements with ambulatory assessment technology were feasible, clinical observations described the voice of patients with depression as low, slow, and hesitant, with these patients speaking in a monotonous and expressionless manner (Kraepelin, 1921; Sundberg, 1998). Voice and speech production may be affected by typical characteristics of the clinical nature of depression; for example, psychomotor retardation, energy loss, and cognitive difficulties also affect the vocal folds, leading to a lower intensity, rate, and loudness of speech, which manifest in a monotone and toneless voice (France et al., 2000; Hashim et al., 2017; Smith et al., 2020). Recent reviews have highlighted the potential of using speech markers to assess a variety of psychiatric disorders (Low et al., 2020), especially depression (Cummins et al., 2015). The use of speech as a marker has several advantages because it can be recorded (1) casually; (2) in a noninvasive manner at people’s homes or in public places (with consent provided); and (3) at low cost because microphones are integrated in many devices such as smartphones, smartwatches, and hearing aids. With the availability of open-source speech analysis software (e.g., open-source Speech and Music Interpretation by Large-Space Extraction (openSMILE) and Praat) and advances in automatic speech processing technologies based on machine learning techniques, research and development on the use of acoustic and linguistic features to identify mood disorders in particular (Low et al., 2020) have been made possible.

Prior Work

Many studies have successfully discriminated between healthy controls and patients with depression based on speech features (Cummins et al., 2015). However, understanding within-person (vs. between-person) depression-related voice changes is essential in detecting new episodes, that is, the secondary prevention. To the best of our knowledge, only a few studies in samples with clinical (not subclinical) depression have examined the variability of speech features within persons (R. Horwitz et al., 2013; Mundt et al., 2007, 2012; Quatieri & Malyska, 2012; Trevino et al., 2011; Yang et al., 2013). In a 6-week treatment-monitoring study, weekly speech samples were obtained from 35 patients with depression using an interactive voice response system (Mundt et al., 2007). Patients with an improvement in depressive symptoms showed a significant increase in pitch and pitch variability, an increase in speech rate, and shorter speech pauses while speaking at their final assessment compared with their baseline assessment. Importantly, patients whose depressive symptoms did not improve did not show these changes.

The data set of Mundt et al. (2007) was reanalyzed multiple times (R. Horwitz et al., 2013; Quatieri & Malyska, 2012; Trevino et al., 2011). Quatieri and Malyska (2012) integrated additional speech features and identified that lower pitch variability, shimmer, and jitter as well as an increased harmonics-to-noise ratio were correlated with lower depression severity. This is in contrast to the study by Mundt et al. (2007), who found that increased pitch variability was associated with lower depression severity, which Quatieri and Malyska (2012) attributed to differences in the set of voice samples analyzed (read speech in the study by Mundt et al., 2007 and free speech in the study by Quatieri & Malyska, 2012).

Trevino et al. (2011) discussed speech rate extraction methods based on the data set of Mundt et al. (2007) and replicated results regarding speech rate in automatically derived phonologically based features. Speech rate was negatively correlated with depression scores and the psychomotor retardation item in particular. Moreover, the authors replicated the finding that speech pauses were positively correlated with depression severity.

Furthermore, Horwitz et al. (2013) reanalyzed a subset of data from the study by Mundt et al. (2007) with a focus on disentangling how speech features relate to the total assessment score and individual symptom items. The authors found a positive correlation between pitch variability and depression scores and a slower speech rate with increasing depression severity. Notably, they analyzed a different speech task and a different depression assessment in comparison with Mundt et al. (2007).

Mundt et al. (2012) replicated their results from Mundt et al. (2007) in a larger study. Here, 105 patients were observed in a 4-week randomized placebo-controlled study. Again, analyses entailed a comparison of the final and baseline assessments. For patients benefiting from the treatment, total pause time was lower, pitch was higher (pitch variability was not assessed), and speech rate was higher. For patients who did not benefit from the treatment, only speech rate increased; however, it increased significantly less than in patients benefiting from the treatment.

Yang et al. (2013) analyzed clinical interviews recorded in 7-week intervals. In contrast to Mundt et al. (2007), they did not find a change in pitch variability with a change in depression severity in the patients but rather in the interviewers. The authors also found shorter switching pauses between patient and interviewer (i.e., both interlocutors) with lower depression severity.

Although not completely consistent, these findings support the assumption that voice features change within individuals when depression severity changes. However, although data were collected at multiple time points during the study (e.g., weekly), except in the study by Yang et al. (2013), the analysis was limited to a comparison between the baseline and final assessments. However, given that the goal is to detect and ultimately prevent new depressive episodes and deterioration, it is essential to understand within-person trajectories of voice features and how they are associated with momentary states with increased granularity. In this study, we used a naturalistic data set where a rapidly acting antidepressant treatment (i.e., sleep deprivation therapy (SDT) (Wirz-Justice & Benedetti, 2020)) was applied to patients experiencing a depressive episode. The antidepressant effect vanishes in most of the cases after recovery sleep. Baseline, the treatment effect of SDT, and relapse can be measured in a matter of four days, making it a preferable setting to study within-person fluctuations.

Aims and Hypotheses

To investigate the within-person relationship between fluctuations in depression severity and fluctuations in speech features, we used a longitudinal data set with an average of 32 (*SD* 19.83) assessments per patient. All patients had experienced an acute depressive episode and undergone SDT (Wirz-Justice & Benedetti, 2020), a chronotherapeutic intervention that can rapidly improve depression symptomatology. The main advantage of this therapeutic is that we maximize the variance of affective states within the data set and ensure sufficient within-person fluctuations over time. As the amount of speech features is immense, resulting in alpha error inflation, we focused on 3 speech features with high face validity that have shown first hints in past research (R. Horwitz et al., 2013; Mundt et al., 2007, 2012; Quatieri & Malyska, 2012; Trevino et al., 2011; Yang et al., 2013). Specifically, we hypothesized that (1) changes in pitch variability, (2) shorter speech pauses, and (3) higher speech rate are associated with lower depression severity. In addition, we assessed the associations of these features with additional momentary affective states (i.e., positive affect, negative affect, valence, energetic arousal, and calmness). We hypothesized that the associations of speech features with negative affect are similar to those for depression severity and that the associations of speech features with the other momentary affective states listed follow the opposite pattern.

Methods

Sample

We used a data set that was collected as part of a pilot study (Sleep Deprivation and Gene Expression (SLEDGE II; German Clinical Trials Register: DRKS00022025) gathering digital phenotypes and multiomics data in a clinical sample undergoing SDT at the Central Institute of Mental Health in Mannheim, Germany. A total of 30 inpatients experiencing acute depressive episodes were enrolled in the study. The patients were diagnosed according to the International Classification of Diseases, Tenth Revision (ICD-10), codes by the senior clinician at admittance to the hospital. All patients received treatment as usual, which also included SDT (for a list of medications, refer to Appendix A2.1). Exclusion criteria were comorbid substance use disorders or personality disorders.

From this sample of 30 patients, the complete data sets of eight patients were excluded from the final analyses (n = 4 did not record any videos; *n* = 1 did not say anything during the videos (23 videos); *n* = 2 had no sound recorded in the videos owing to technical issues (30 videos); and *n* = 1 recorded only two videos); thus, the final sample consisted of 22 patients (*n* = 12, 55% male) aged between 18 and 63 (mean 33.5, *SD* 12.4; median 29, IQR 23.25-42.75) years.

Ethical Considerations

The study was approved by the Ethics Committee II of the Medical Faculty Mannheim, University of Heidelberg (2013-563N-MA). All patients received detailed information about the aims and procedures of the study and provided informed consent. Patients could withdraw from the study at any time and did not receive any compensation for participation. Data was deidentified to ensure privacy.

Study Procedure

Patients were given a study smartphone (Nokia 4.2 or Samsung Galaxy J7) at the beginning of the study (day 0), instructed on how to use it, and (if necessary) performed test runs supervised by the study personnel. A telephone number for technical support and an information sheet regarding the ambulatory assessment procedure were handed out. Data were collected using movisensXS software (https://movisens.com/en). Patients underwent SDT as part of their depression treatment, which involves staying awake for approximately 36 hours. Treatment effect and relapse can be measured in a matter of 4 days (Wirz-Justice & Benedetti, 2020), thus ensuring a maximum of within-person variance in the data set. After at least one day of baseline assessment (day 0), SDT was conducted on day 1. Patients stayed awake from 6 AM on day 1 to 6 PM on day 2. Recovery sleep was allowed from 6 PM on day 2 until 1 AM on day 3. Data were collected before, during, and after SDT for up to 26 days. In the first week of the study, smartphones sent prompts three times per day (morning, afternoon, and evening); in addition, self-initiated assessments were possible to report specific events or to catch up with missed assessments. To reduce the burden on patients, the sampling schema was altered to two prompts per day (morning and evening). With each prompt, patients were requested to fill out items concerning their affective state and to record a selfie video reporting how they felt currently. Patients returned the smartphone at the end of the study. The study personnel uploaded the data from the smartphones to the movisensXS platform (https://movisens.com/en/) and then downloaded the data for analysis.

Ambulatory Assessment: E-diary Ratings and Selfie Videos

The data set contains three sets of momentary assessments in the form of e-diary ratings at each prompt (Appendix A2.2-A2.4): (1) the short version of the Allgemeine Depressionsskala (ADS-K; Hautzinger, 1988) adapted to momentary assessment with 14 items on depressive mood rated on a scale ranging from 0 = *rarely* to 3 = *mostly* (we left out the item regarding sleep from the original questionnaire because its inclusion was not reasonable in the momentary assessment design); (2) a total of 15 positive (cheerful, content, energetic, enthusiastic, relaxed, and happy) and negative (lonely, sad, insecure, anxious, depressed, low-spirited, guilty, distrustful, and irritable) affect items (Myin-Germeys et al., 2003) rated on a 5-point Likert scale ranging from 1 = *not at all* to 5 = *very much*; and (3) a 6-item short version of the Multidimensional Mood Questionnaire (MDMQ; Wilhelm & Schoebi, 2007) capturing time-varying momentary fluctuations in daily life on the affect dimensions of valence (*unwell* to *well* and *discontent* to *content*), energetic arousal (*without energy* to *full of energy* and *tired* to *awake*), and calmness (*tense* to *relaxed* and *agitated* to *calm*). The items were presented on visual analog scales with two poles and a slider from 0 to 100. For each of the constructs, we computed mean values per scale, resulting in six outcome variables (depressive symptoms, positive affect, negative affect, valence, energetic arousal, and calmness). For the ADS-K, we also report sum scores as described in the tool’s manual; however, to increase comparability among outcomes, we used the mean value for analyses. If necessary, we recoded items such that higher values indicated a (1) higher intensity of depressive symptoms, (2) higher positive affect, (3) higher negative affect, (4) higher positive valence, (5) higher energetic arousal, and (6) higher calmness.

In addition to the aforementioned e-diary ratings, patients were requested to record selfie videos with the following instructions: *Please keep the camera stable during the recording and record your whole face. Please describe in 10-20 seconds how you currently feel.*

Clinical Assessments

The Montgomery–Åsberg Depression Rating Scale (MADRS; Montgomery & Åsberg, 1979) was completed in the morning at four time points (baseline, morning before sleep deprivation, one week after sleep deprivation, and two weeks after sleep deprivation) and once at midday (the day after sleep deprivation night). The MADRS is a 10-item expert assessment of depressive symptom severity over the past week, with items rated on a 7-point scale ranging from 0 to 6; higher scores indicate higher severity.

Data Preprocessing

The data set contained 899 selfie videos in mp4 format. The full set of videos of four of the 30 patients had to be excluded owing to the reasons mentioned previously (55 videos) and additional two videos had to be excluded because of technical damage (2 videos). As our research questions focused on audio data (not visual data), we extracted the audio tracks of the remaining 842 from the original 899 selfie videos using the *ffmpeg* package in Python and archived them as wav files (sampling rate: 48 kHz; mono=1 channel). We excluded test runs (14 videos), accidental short recordings with no content (29 videos), recordings during which the microphone was masked by the patient 27 videos), and assessments in which one of the two corresponding assessments (speech or affective state) was missing (18 videos). In addition, if two consecutive assessments were <15 minutes apart from each other, only the first assessment was kept unless its audio quality was insufficient or only the second assessment included assessments of affective states; in such cases, the second assessment was kept (21 videos). We also excluded recordings with background noise that restricted speech intelligibility (9 videos) or that included the speech of third parties (8 videos). We filtered the remaining 716 recordings using *DeepFilterNet2* (Schröter et al., 2022) to remove background noise.

Acoustic Features

For our main analyses, we focused on the acoustic features pitch variability, speech pauses, and speech rate (Table 1). We restricted the number of features to limit α error inflation and selected specifically these three features because they revealed sufficient empirical support to warrant an explicit hypothesis. We extracted acoustic features of the final recordings (n = 716) using the *functionals (v02)* of the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS; Eyben et al., 2016) of the open-source toolkit *openSMILE* (Eyben et al., 2010) implemented in Python (https://github.com/audeering/opensmile-python). eGeMAPS is a minimalistic set of acoustic features recommended for clinical speech analysis; it helps to guarantee comparability between studies, given the proliferation of speech features. Features related

Table 1

Overview of Extracted Speech Features

Speech feature	Technical feature	Explanation
Pitch variability	F0semitoneFrom27.5Hz_sma3nz_stddevNorm	SD of the F0 perceived as the extent to which a person's pitch changes (in Hz)
Speech pauses	MeanUnvoicedSegmentLength	Mean of the length of unvoiced regions approximating silent parts of the speech sample (in seconds)
Speech rate	Words per second	Ratio of words counted on the basis of the automatically transcribed and manually corrected text divided by the duration of the speech sample

Note. F0 = fundamental frequency.

to frequency, energy, spectrum, and tempo are included in the set. Pitch variability is represented by the *SD* of the logarithmic fundamental frequency (F0) on a semitone frequency scale starting at 27.5 Hz and measured in hertz. F0 is the lowest frequency of the speech signal and is perceived as pitch. Speech pauses are approximated as the mean length of unvoiced regions (F0=0) measured in seconds. With respect to speech rate, a transcription of the recordings is necessary, which we obtained using an automatic speech recognition system according to published procedures (Abulimiti et al., 2020). We corrected the transcripts manually. To determine speech rate, we calculated the ratio of words divided by the duration of the voice sample.

Beside our main analyses based on pitch variability, speech pauses, and speech rate, we decided to integrate further eGeMAPS features in an exploratory analysis. These features have been recommended in the context of affective states in particular because they contain additional cepstral and dynamic features (Eyben et al., 2016). We included the following features in the exploratory analyses: for voiced and unvoiced regions together, the mean and *SD* of the mel-frequency cepstral coefficients (MFCCs) 1 to 4 and spectral flux difference of the spectra of two consecutive frames; for voiced regions, the formant 2 to 3 bandwidths along with spectral flux and MFCCs 1 to 4; and for unvoiced regions, the mean and *SD* of the spectral flux (Eyben et al., 2016).

Statistical Analysis

In addition to the *mean, SD,* and *range*, we present *min* and *max* as the mean of all patients' minimum and maximum scores, respectively, of each parameter throughout the whole study. Moreover, following the recommendations by Snijders and Bosker (2011), we computed Pearson correlation analyses with person-mean–centered variables to evaluate the relationship between affective scores and speech features. To generate person-mean–centered variables, we subtracted the individual's mean from their score, which represents the variation around the individual's mean.

To evaluate psychometric properties, we calculated McDonald ω as the reliability coefficient using the *multilevelTools* package in R. For the MDMQ subscales, we used the *misty* package in *R* to calculate the Spearman-Brown corrected correlation coefficients because the subscales consist of only two items (Eisinga et al., 2013). For the MADRS score at the time of inclusion, we calculated Cronbach α using the *psych* package in R.

To analyze the within-person association of speech features and subjectively evaluated affective states, we used multilevel modeling (Snijders & Bosker, 2011) using the *nlme* package in R. Multilevel modeling offers two specific advantages for the given data: (1) separation of within-person effects from between-person effects and (2) allowing and considering different numbers of assessments per patient. Before the analyses, we centered time-variant level-1 predictors (pitch variability, speech pauses, and speech rate) at the person level and included the predictors time and time² in minutes (each centered at 2 PM) as covariates. To facilitate the comparison of the magnitude of effects among different predictors, we report standardized beta coefficients (standardized β) according to the recommendations by Hox and van de Schoot (2013) following the equation: standardized $\beta = \beta \times (SD_{\text{predictor}} / SD_{\text{outcome}})$. We further calculated R^2_{Hox} values according to the recommendation by Maas and Hox (2005) following the equation: $R^2_{\text{Hox}} = (\sigma^2_{\text{null}} - \sigma^2_{\text{model}}) / \sigma^2_{\text{null}}$. We set the α level at 5% and applied Bonferroni corrections for exploratory analyses (α_{adj}=.002). We performed all analyses in *R* (version 4.2.1, 2022-06-23).

Our analyses can be split into four parts: the calculation of intraclass correlation coefficients (ICCs); separate models with all speech features as predictors and all affective scores as outcomes; combined models with all speech features as simultaneous predictors; and exploratory analyses, including additional speech features. Specifically, we first descriptively investigated whether our study procedure resulted in sufficient within-person variance. For this purpose, we calculated ICCs, including all momentary affective ratings and speech recordings, regardless of whether they were assessed before, during, or after SDT. In general, the ICC indicates the amount of between-person variance in unconditional (null) models. The 2-level models analyzed contained repeated measures (level 1) that were nested within patients (level 2). The second step contained our main analysis: we calculated separate models for each speech feature (pitch variability (model set 1), speech pauses (model set 2), and speech rate (model set 3)) and each affective state (depression severity (ADS-K), positive affect, negative affect, valence, energetic arousal, and calmness), resulting in 18 models. In the third step, to evaluate the relative significance of pitch variability, speech pauses, and speech rate, we constructed combined models for each of the affective scores, including all three features simultaneously (six models). In the fourth step, exploratory analyses were conducted with the inclusion of 24 additional speech features from eGeMAPS (Appendix A2.5). These features were used as predictors for each of the affective scores separately.

Results

Descriptive Statistics

We included 716 speech-state pairs (mean 32, *SD* 19.83 per patient) in the final analysis. The mean MADRS score at the time of inclusion assessment was 30.1 (*SD* 5.8). This corresponds to 18 patients with moderate depression and four patients with severe depression out of 22 patients at study inclusion.

Regarding depressive symptoms (ADS-K; scale 0-3), patients had a mean score of 1.2 (*SD* 0.6; *min* 0.7, *max* 2.0) and a mean sum score of 16.9 (*SD* 8.1; *min* 9.6, *max* 26.1). At inclusion, the mean ADS-K score was 1.4 (*SD* 0.6; range 0.4-2.8), and the mean sum score was 20.0 (*SD* 8.4; range 6-39). For positive and negative affect (scale 1-5), the mean scores were 2.1 (*SD* 0.8; *min* 1.3, *max* 3.1) and 2.3 (*SD* 1.0; *min* 1.4, *max* 3.9), respectively; on the MDMQ (scale 1-100) valence subscale, the mean score was 44.9 (*SD* 21.5; *min* 9.4, *max* 67.5); on the energetic arousal subscale, the mean score was 41.7 (*SD* 21.0; *min* 16.4, *max* 62.7); and on the calmness subscale, the mean score was 43.8 (*SD* 22.8; *min* 6.9, *max* 70.7). The ICCs were 0.47 for the ADS-K, 0.45 for positive affect, 0.59 for negative affect, 0.27 for energetic arousal, 0.25 for valence and 0.40 for calmness, that is, the following amount of variance in the momentary assessments can be attributed to within-person fluctuations: 53% for the ADS-K, 55% for positive affect, 41% for negative affect, 73% for energetic arousal, 75% for valence, and 60% for calmness.

Regarding speech features, the mean pitch variability was 0.32 Hz (*SD* 0.09; *min* 0.14, *max* 0.44), the mean speech pause length was 0.26 seconds (*SD* 0.12; *min* 0.17, *max*

0.47), and the mean speech rate was 1.77 words per second (*SD* 0.57; *min* 1.16, *max* 2.75). The ICCs were 0.66 for pitch variability, 0.36 for speech pauses, and 0.57 for speech rate. This corresponds to the following amount of variance in the speech feature assessments that can be attributed to within-person fluctuations: 34% for pitch variability, 64% for speech pauses, and 43% for speech rate.

Correlational analyses (see Appendix A2.6) included between 698 and 716 observations depending upon the specific pairing. We found correlations among and between affective scores and speech features, except for pitch variability and speech rate, neither of which correlated with negative affect and calmness; in addition, there was no correlation between pitch variability and speech rate. Specifically, ADS-K scores correlated negatively with positive affect, all MDMQ subscales, and speech rate and correlated positively with negative affect, pitch variability, and speech pauses. Negative affect showed the same pattern, except for the pairings with pitch variability and speech rate, for which no correlations were found. Regarding positive affect, we found the opposite correlation pattern, that is, positive correlations with all MDMQ subscales and speech rate and negative correlations with pitch variability and speech pauses. The MDMQ subscales showed the same relationships as positive affect, except for the pairing between calmness and pitch variability and speech rate, for which no correlations were found. Within speech features, we found a negative correlation between pitch variability and speech pauses, no correlation between pitch variability and speech rate, and a negative correlation between speech pauses and speech rate. Overall, correlations among affective scores were strong (r > .5). Correlations among speech features as well as between affective scores and speech features were weak (r < .2), except for a strong negative correlation between speech pauses and speech rate.

The psychometric properties for momentary affective ratings were good to excellent. Specifically, McDonald ω values (Geldhof et al., 2014) were 0.87 (within-person) and 0.90 (between-person) for depressive symptoms (ADS-K), 0.87 (within-person) and 0.95 (between-person) for positive affect, and 0.87 (within-person) and 0.96 (between-person) for negative affect. The Spearman-Brown coefficients were 0.83 (within-person) and 0.94 (between-person) for valence, 0.74 (within-person) and 0.89 (between-person) for energetic arousal, and 0.74 (within-person) and 0.89 (between-person) for calmness. Cronbach α for the MADRS score at the time of inclusion was acceptable (.67).

Association Between Speech Features and Momentary Affective Scores

In Tables 2 and 3, we present the fixed effects of pitch variability, speech pauses, and speech rate separately for each affective state. Details, including the effects of time and time^2, are presented in Appendix A2.7.

ADS-K Scores

In the column entitled ADS-K (Table 2), we report the results of all models with ADS-K scores as the outcome. Pitch variability (standardized β = .14; *p* = .007), speech pauses (standardized β = .10; *p* = .005), and speech rate (standardized β = −.10; *p* = .02) were significantly associated with the ADS-K score, indicating that higher pitch variability, longer speech pauses, and lower speech rate are associated with more severe depressive symptomatology.

Table 2

Multilevel Linear Regression Analysis to Predict Depression and Positive and Negative Affect: Fixed Effects for Pitch Variability, Speech Pauses, and Speech Rate

Predictors	Outcome														
	ADS-K					Positive affect					Negative affect				
	β	Stand. β	*SE*	R^2_{Hox} (%)	*p*	β	Stand. β	*SE*	R^2_{Hox} (%)	*p*	β	Stand. β	*SE*	R^2_{Hox} (%)	*p*
Model set 1															
Intercept	127	-	0.10	-	<.001	2.10	-	0.13	-	<.001	2.45	-	0.16	-	<.001
Pitch variability	.88	.14	0.32	1	**.007**	-1.50	-.18	0.42	1	**<.001**	.85	.08	0.43	1	.05
Model set 2															
Intercept	127	-	0.10	-	<.001	2.09	-	0.13	-	<.001	2.46	-	0.16	-	<.001
Speech pauses	.52	.10	0.18	1	**.005**	-1.16	-.17	0.24	17	**<.001**	.76	.09	0.25	2	**.002**
Model set 3															
Intercept	127	-	0.10	-	<.001	2.10	-	0.13	-	<.001	2.45	-	0.16	-	<.001
Speech rate	-.11	-.10	0.05	<1	**.02**	.26	.18	0.06	2	**<.001**	-.13	-.08	0.07	1	**.04**

Note. ADS-K = Allgemeine Depressionsskala Kurzform. Stand. β = standardized β coefficient. Statistical significance printed in bold.

Table 3

Multilevel Linear Regression Analysis to Predict Valence, Energetic Arousal, and Calmness: Fixed Effects for Pitch Variability, Speech Pauses, and Speech Rate

Predictors	Outcome														
	Valence					Energetic arousal					Calmness				
	β	Stand. β	*SE*	R^2_{Hox} (%)	*p*	β	Stand. β	*SE*	R^2_{Hox} (%)	*p*	β	Stand. β	*SE*	R^2_{Hox} (%)	*p*
Model set 1															
Intercept	43.72	-	2.70	-	<.001	42.82	-	2.71	-	<.001	40.97	-	3.39	-	<.001
Pitch variability	-36.50	-.16	13.61	1	**.008**	-33.21	-.15	12.48	1	**<.001**	-11.52	-.05	12.82	<1	.37
Model set 2															
Intercept	43.26	-	2.69	-	<.001	42.71	-	2.71	-	<.001	40.58	-	3.39	-	<.001
Speech pauses	-34.06	-.19	7.71	3	**<.001**	-14.06	-.08	7.14	1	**.049**	-24.27	-.12	7.27	5	**<.001**
Model set 3															
Intercept	43.56	-	2.70	-	<.001	42.77	-	2.71	-	<.001	40.86	-	3.39	-	<.001
Speech rate	6.49	.17	2.03	2	**.001**	4.13	.11	1.87	1	**.03**	3.43	.09	1.91	5	.07

Note. Stand. β = standardized β coefficient. Statistical significance printed in bold.

Positive and Negative Affect

In the columns entitled Positive affect and Negative affect (Table 2), we show results for positive affect and negative affect, respectively, as outcomes. Pitch variability (standardized β = −.18; p < .001), speech pauses (standardized β = −.17; p < .001), and speech rate (standardized β = .18; p < .001) were significantly associated with positive affect, indicating that lower pitch variability, shorter speech pauses, and higher speech rate are associated with higher positive affect. The associations between negative affect and speech features were in the opposite direction of the associations between positive affect and the speech features just presented: speech pauses (standardized β = .09; p = .002) and speech rate (standardized β = −.08; p = .04) were significantly associated with negative affect, indicating that longer speech pauses and lower speech rate are associated with higher negative affect. We further found a trend with respect to the association between pitch variability and negative affect, but this result was not statistically significant (standardized β = .08; p = .05). In addition, we found trends with respect to the associations between negative affect and time and negative affect and time^2, specifically in the models that included pitch variability (time: standardized β = .04; p = .08), speech pauses (time: standardized β = .04; p = .08; time^2: standardized β < .01; p = .06), and speech rate (time: standardized β = .04; p = .09), but these results were not statistically significant.

MDMQ Results

In the columns entitled Valence, Energetic arousal, and Calmness (Table 3), we present the results for the MDMQ. *Pitch variability* (standardized β = −.16; p = .008), *speech pauses* (standardized β = −.19; p < .001), and *speech rate* (standardized β = .17; p = .001) were significantly associated with valence, indicating that lower *pitch variability*, shorter *speech pauses*, and higher *speech rate* are associated with higher (i.e., positive) valence. In the model that included valence and *speech pauses*, we found a significant association between time^2 and valence (standardized β < .001; *p* = .03). In addition, we found trends with respect to the associations between valence and time^2, specifically in the models that included *pitch variability* (time: standardized β < .01; *p* = .098) and *speech rate* (time: standardized β < .01; *p* = .07), but these results were not statistically significant. Moreover, *pitch variability* (standardized β = −.15; *p* < .001), *speech pauses* (standardized β = −.08; *p* = .049), and *speech rate* (standardized β = .11; *p* = .03) were significantly associated with energetic arousal, indicating that lower *pitch variability*, shorter speech paus*es*, and higher *speech rate* are associated with higher energetic arousal. In all model combinations of energetic arousal and each speech feature, we found significant associations between time and energetic arousal (standardized β = −.11; *p* < .001) and time^2 and energetic arousal (standardized β < .01; *p* < .001). Furthermore, *speech pauses* (standardized β = −.12; *p* < .001) were significantly associated with calmness, indicating that shorter *speech pauses* are associated with greater calmness. In all model combinations of calmness and each speech feature, we found significant associations between time^2 and calmness (standardized β < .01; *p* = .013 for *pitch variability*, *p* = .003 for *speech pauses*; *p* = .009 for *speech rate*). In addition, we found a trend with respect to the association between *speech rate* and calmness (standardized β = .09; *p* = .07), but this result was not statistically significant.

Combined Models

In Table 4 and Table 5 we display the results for the combined models that included all three speech features. In the model of ADS-K scores, associations with pitch variability (standardized β = .17; *p* <.001) and speech pauses (standardized β = .12; *p* = .01) remained statistically significant. Regarding positive affect, associations with pitch variability (standardized β = −.23; p < .001) and speech pauses (standardized β = −.19; p < .001) remained statistically significant. We further found a trend regarding the association between positive affect and time (standardized β = −.05; p =.09), but this result was not statistically significant. Regarding negative affect, associations with pitch variability (standardized β = .12; p = .008), speech pauses (standardized β = .12; p = .005), time (standardized β = .05; p = .03), and time^2 (standardized β < .01; p = .03) remained statistically significant. In the model of valence, associations with pitch variability (standardized β = −.22; p < .001), speech pauses (standardized β = .22; p < .001), and time^2 (standardized β < .01; p = .01) remained statistically significant. Regarding energetic arousal, associations with pitch variability (standardized β = −.17; p = .003), time (standardized β = .12; p < .001), and time^2 (standardized β < .01; p < .001) remained statistically significant. Regarding calmness, associations with speech pauses (standardized β= − .17; p = .002) and time^2 (standardized β < .01; p = .002) remained statistically significant. We further found a trend for the association between calmness and pitch variability (standardized β = .09; p = .097), but this result was not statistically significant.

Exploratory Analysis

Analyzing additional speech features, we found significant associations of the equivalent sound level, the mean of spectral flux, and the mean of spectral flux of voiced regions only, individually, with all affective scores (Appendix A2.8). With respect to equivalent sound level, this indicates that louder voice samples were linked to improved affective states (ADS-K: standardized β = −.30; positive affect: standardized β = .34; negative affect: standardized β = −.21; valence: standardized β = .29; energetic arousal: standardized β = .26; and calmness: standardized β = .19); with respect to the mean of

Table 4

Multilevel Linear Regression Analysis to Predict Momentary Depression, Positive Affect, and Negative Affect: Fixed Effects for the Combined Models

Predictors	Outcomes[a]											
	ADS-K				Positive Affect				Negative Affect			
	β	Stand. β	*SE*	*p*	β	Stand. β	*SE*	*p*	β	Stand. β	*SE*	*p*
Intercept	1.28	-	0.10	<.001	2.08	-	0.13	<.001	2.47	-	0.16	<.001
Time	<0.01	.02	<0.01	.42	<-0.01	-.05	<0.01	.09	<0.01	.05	<0.01	**.03**
Time^2	<0.01	<.001	<0.01	.44	<0.01	<.001	<0.01	.31	<0.01	<.01	<0.01	**.03**
Pitch variability	1.11	.17	0.33	**<.001**	-1.96	-.23	0.43	**<.001**	1.19	.12	0.45	**.008**
Speech pauses	.64	.12	0.26	**.01**	-1.29	-.19	0.33	**<.001**	0.99	.12	0.35	**.005**
Speech rate	<-0.01	<.001	0.07	.99	0.04	.03	0.09	.66	0.04	.02	0.09	.68

Note. Stand. β = standardized β coefficient. [a]R^2_{Hox} for ADS-K = 2%, for positive affect = 6%, and for negative affect = 2%. Statistical significance printed in bold.

Table 5

Multilevel Linear Regression Analysis to Predict Momentary Valence, Energetic Arousal, and Calmness: Fixed Effects of the Combined Models

Predictors	Outcomes[a]											
	Valence				Energetic arousal				Calmness			
	β	Stand. β	*SE*	*p*	β	Stand. β	*SE*	*p*	β	Stand. β	*SE*	*p*
Intercept	42.95	-	2.68	<.001	42.48	-	2.71	<.001	40.45	-	3.38	<.001
Time	<0.01	.03	<0.01	.48	<-0.01	.12	<0.01	**<.001**	<-0.01	.01	<0.01	.89
Time²	<0.01	<.01	<0.01	**.01**	<0.01	<.01	<0.01	**<.001**	<0.01	<.01	<0.01	**.002**
Pitch variability	-49.01	-.22	13.76	**<.001**	-37.74	-.17	12.78	**.003**	-21.75	.09	13.07	.097
Speech pauses	-41.01	.22	10.73	**<.001**	-12.97	.07	9.97	.19	-32.53	-.17	10.20	**.002**
Speech rate	-0.64	.02	2.76	.82	1.96	.05	2.56	.44	-2.28	.06	2.62	.38

Note. Stand. β = standardized β coefficient. [a]R^2_{Hox} for valence = 4%, for energetic arousal = 5%, and for calmness= 2%. Statistical signify-cance printed in bold.

spectral flux, this indicates that a faster change in the spectrum was linked to better affective states (ADS-K: standardized β = −.22, positive affect: standardized β = .28, negative affect: standardized β = −.15, valence: standardized β = .21, energetic arousal: standardized β = .17, and calmness: standardized β = .27); and with respect to the mean of spectral flux of voiced regions only, this indicates that a faster change in the spectrum in voiced regions was linked to better affective states (ADS-K: standardized β = −.23, positive affect: standardized β = .28, negative affect: standardized β = −.15, valence: standardized β = .20, energetic arousal: standardized β = .20, and calmness: standardized β = .16). Regarding the additional speech features, the following significant associations were found: the mean of spectral flux of unvoiced regions only was associated with positive affect, indicating that a faster change in the spectrum in unvoiced regions was linked to improved positive affect (standardized β = .13); and the mean of the MFCC 2 of voiced regions only was significantly associated with energetic arousal, indicating that a higher mean was linked to lower energetic arousal (standardized β = −.15). Furthermore, we revealed a significant association between the *SD* of the MFCC 4 of voiced regions only ADS-K scores (standardized β = .13) as well as positive affect (standardized β = −.10) and negative affect (standardized β = .09). Specifically, smaller SDs were linked to higher positive affect, reduced negative affect, and lower ADS-K scores.

Discussion

Principal Findings

This is the first study to investigate whether speech features are associated with depression severity and momentary affective states in a longitudinal data set of patients with a depressive episode undergoing SDT. Our findings showed that lower pitch variability, higher speech rate, and shorter speech pauses were associated with better momentary states (i.e., lower depression severity; higher positive affect and lower negative affect; and higher positive valence, energetic arousal, and calmness), supporting prior clinical observations with innovative methods applied to an intensive longitudinal data set.

Lower depression severity was accompanied by shorter speech pauses. This is in line with past research findings reporting that shorter speech pauses were associated with lower depression severity (Mundt et al., 2007, 2012; Trevino et al., 2011; Yang et al., 2013). Our findings extend prior results because we also found an association between speech pauses and affective states more broadly, not limited to depressed mood. Regarding speech rate, we revealed associations with depression severity and all other affective state scales except for calmness. In particular, we found that higher speech rate was associated with lower depression symptomatology and lower negative affect, higher positive affect, higher positive valence, and higher energetic arousal. This is in line with prior research (R. Horwitz et al., 2013; Mundt et al., 2007, 2012; Trevino et al., 2011), in which a higher speech rate was found for patients who benefited from treatment.

Regarding pitch variability, we found support for our hypothesis that pitch variability changes with depression severity; more precisely, lower pitch variability was associated with lower depression symptomatology. This is in line with the studies by Quatieri and Malyska (2012) and Horwitz et al. (2013), where a positive correlation between pitch variability and depression severity was found. However, the results reported in the studies by Mundt et al. (2007) and Yang et al. (2013) contrasted with ours and those found in the studies by Quatieri and Malyska (2012) and Horwitz et al. (2013), that is, that higher pitch variability was associated with lower depression severity. A possible explanation for contradictory results in major depression are the heterogeneity of (1) the depression phenotype per se because diagnosis criteria include >400 possible symptom combinations (Goldberg, 2011; Østergaard et al., 2011); and (2) the questionnaires, assessment approaches, statistical analyses, and speech feature extraction tools used in these studies. The within-person research design approach underlying our data set addressed the heterogeneity of the depression phenotype at least partially. Furthermore, we analyzed free speech collected naturally in a selfie task, whereas in the study by Mundt et al. (2007), read speech was used in the analyses. In line with what is suggested in the study by Quatieri and Malyska (2012), this could also be a reason for the contradictory results. However, because assessing within-person fluctuations in daily life increases ecological validity, we regard our results as an important contribution.

Observing the full picture of associations, we note that the results for all three speech features are similar and do not provide evidence of specific associations (e.g., association of one specific speech feature with one specific momentary affective state), showing no distinct patterns of momentary states for each speech feature. This is reasonable because the constructs overlap in content (e.g., patients experiencing depression experience higher negative affect and lower positive affect).

In terms of the combined models evaluating the relative importance of the features, we found that in the four models (ADS-K, valence, positive affect, and negative affect) both pitch variability and speech pauses remained significant, whereas speech rate did not. Pitch variability remained the only significant parameter in the model of energetic arousal, and speech pauses remained the only significant parameter in the model of calmness. This suggests that pitch variability and speech pauses are speech features rather independent of each other, whereas the high correlation between speech pauses and speech rate might account for the fact that only one of these features (in this case, speech pauses) remained a significant predictor.

Limitations

First, this study examined a limited set of three speech features. Instead of applying brute force methods involving thousands of technical speech features, we selected speech features based on previous work and with high face validity, restricting the scope of our analysis. Although we did expand our scope of features in the exploratory analysis, it is very likely that other configurations and features (e.g., the *ComParE* feature set containing 6373 features; Schuller et al., 2013) might also be predictive of affective states. Future work is needed to compare theory-driven approaches with brute force data-driven machine learning methods to find the best possible combination of speech features also considering aspects of computational power. However, selecting the features on a theoretical basis and restricting their pure number limits alpha error inflation and should increase replicability.

Second, although the sample size of our study was limited, this was a true within-person design with many data points per patient. In addition, we regard this study as a pilot study providing important indications regarding feasibility in a clinical context. As some patients dropped out of the study, and some recordings had to be excluded, in future studies, data collection needs to be integrated better into clinical routines. Moreover, the instructions for patients may need to be revised to reduce the likelihood of missing data and recording errors. However, the data set at hand is still unique in the relatively high number of assessments per patient and the applied SDT, which yielded meaningful variation in the depression severity within a short time period. From a theoretical perspective, it is crucial to emphasize that to uncover existing relations among variables, meaningful variance in both parameters is needed.

Third and last, selfie videos were recorded in a clinical environment, which may limit generalizability to other contexts. In future studies, ambulant patients could be integrated and other environments explored to evaluate the replicability of the results. However, our approach, which involved sampling free speech, offers higher ecological validity to reading standardized text paragraphs because it provides a closer representation of people's everyday lives. The development of passive sensing will be helpful in this context (i.e., the random assessment of audio bits in an ecological environment). To date, automated passive voice recordings in nonprotected environments have been restricted in 2-party consent states, such as Germany. However, in single-party consent states, a few speech-related applications can be used in the wild (e.g., the Electronically Activated Recorder; Mehl, 2006). Although the development of technical devices is ongoing, future studies will have to consider ethical issues related to voice recording in natural settings (e.g., ensuring that no third parties who did not give informed consent are recorded).

Conclusions

Our study provides evidence that fluctuations in the speech features pitch variability, speech pauses, and speech rate are associated with fluctuations in depression severity and other momentary affect states. Notably, the data were collected from clinically diagnosed patients (no subclinical sample or staged emotions) experiencing an acute depressive episode. A particularly important advantage is that our longitudinal ambulatory assessment data set ensured a maximum of within-person dynamics of depressive parameters within a short time period by applying a sleep deprivation intervention design. This is of great importance because future technology will try to predict upcoming depressive episodes on an individual level and will need information on within-person trajectories. For the development of such tailored precision medicine tools, pitch variability, speech pauses, and speech rate present promising features. Our research is a step forward on the path to developing an automated depression monitoring system, facilitating individually tailored treatments and increased patient empowerment.

CHAPTER 3

ARTICLE 2:

SPEECH-BASED MACHINE LEARNING FOR PREDICTING MOMENTARY DEPRESSION SEVERITY

This chapter is based on a manuscript that has been submitted to *JMIR Mental Health* as

Hartnagel, L.-M., Emden, D., Foo, J. C., Streit, F., Witt, S. H., Frank, J., Limberger, M. F., Schmitz, S., Gilles, M., Rietschel, M., Hahn, T., Ebner-Priemer, U. W., & Sirignano, L. [under review]. Speech-based Machine Learning for Momentary Depression-Severity Prediction in Acutely Depressed Patients undergoing Sleep Deprivation Therapy

Abstract

Background: Mobile devices for remote monitoring are inevitable tools to support treatment and patient care, especially in recurrent diseases such as Major Depressive Disorder. The aim of this study was to learn if machine learning (ML) models based on longitudinal speech data are helpful in predicting momentary depression severity. Data analyses were based on a dataset including 30 inpatients during an acute depressive episode receiving Sleep Deprivation Therapy in stationary care, an intervention inducing a rapid change in depressive symptomatology in a relatively short period of time. Using an ambulatory assessment approach, we captured speech samples and assessed concomitant depression severity via self-report questionnaire over the course of three weeks (before, during, and after therapy). We extracted 89 speech features from the speech samples using the eGeMAPS parameter set from openSMILE and the additional parameter speech rate.

Objective: We aimed to understand if a multi-parameter ML approach would significantly improve the prediction compared to previous statistical analyses, and, in addition, which mechanism for splitting training and test data was most successful, especially focusing on the idea of personalized prediction.

Methods: To do so, we trained and evaluated a set of > 500 ML pipelines including random forest, linear regression, support vector regression, and eXtreme gradient boosting regression models and tested them on five different train-test split scenarios: a group 5-fold nested cross-validation on subject level, a leave-one-subject-out approach, a chronological split, an odd-even split, and a random split.

Results: In the 5-fold cross-validation, the leave-one-subject-out, and the chronological split approaches, none of the models were statistically different from random chance. The other two approaches produced significant results for at least one of the models tested, with similar performance. In total, the superior model was an XGBoost regression in the odd-even split approach (R^2 = 0.339, MAE = 0.38; both p < .001), indicating that 33.9% of the variance in depression severity could be predicted by the speech features.

Conclusion: Overall, our analyses highlight that ML fails to predict depression scores of unseen patients, but prediction performance increased strongly compared to our previous analyses with multilevel models. We conclude that future personalized ML models might improve prediction performance even more, leading to better patient management and care.

Introduction

Major depressive disorder (MDD) is a major global public health challenge imposing a substantial burden on individuals and society as a whole (Vos et al., 2020). Due to the recurrent nature of MDD in many patients, relapse prevention is an important treatment goal (Benasi et al., 2021). Longitudinal symptom monitoring is crucial, especially for relapse prevention (Benasi et al., 2021), as mood deterioration and prodromal symptoms can be detected in time and additional treatment can be initiated before a severe episode fully develops. However, traditional retrospective symptom questionnaires and classification interviews typically consider the last two weeks of symptomatology (Colombo et al., 2019), which might not be useful for the rapid detection of impending prodromal symptoms. More specifically, even an unrealistic scenario of conducting classification interviews every two weeks might delay the detection of a new episode by weeks (Ebrahimi et al., 2021; Fried et al., 2022). Accordingly, approaches are needed that operate at a higher frequency, enabling to detect prodromal symptoms e.g. on a daily basis.

Leveraging on smartphone-based data collection, promising avenues are being opened to support the traditional monitoring of MDD symptoms (Abd-Alrazaq et al., 2023; Torous et al., 2016). Offering continuous, unobtrusive, near real-time, active and passive everyday life data collection, the use of ambulatory assessment (AA) increases ecologically valid insights into the lives of people living with mental disorders (Ebner-Priemer et al., 2020; Trull & Ebner-Priemer, 2014). Widespread personal digital devices such as smartphones are used to capture momentary self-reported symptoms and behaviors as patients go about their normal daily activities in their natural environment (Ebner-Priemer & Trull, 2009). As clear biomarkers for MDD are lacking (Rimti et al., 2023), the identification of behavioral markers that can be objectively derived from digitally captured everyday life behavior has great potential to increase automated detection of new episodes, ultimately improving depression care (Abd-Alrazaq et al., 2023; De Angel et al., 2022; Zarate et al., 2022).

Speech has been discussed as one such potential behavioral marker (Low et al., 2020). As early as 1921, Emil Kraeplin observed that patients with MDD tended to speak with a lower speech rate, more monotonously, and at a lower pitch compared to healthy individuals (Kraepelin, 1921). Since then, many studies have described further depression-related altered speech characteristics (Cummins et al., 2015; Low et al., 2020). However, the research field faces several challenges such as the sheer limitless volume of potential speech features. Inference statistics require a theory-driven selection of parameters, as combining thousands of them increases the alpha error (Wadle & Ebner-Priemer, 2023). Machine learning (ML) techniques offer a data-driven alternative, allowing a variety of parameters to be explored without the need for a priori parameter restriction.

Most studies investigating speech in MDD (independent of using ML or classical inferential statistics) use case-control designs, comparing speech samples (or often a single sample) of patients with MDD to healthy controls (Low et al., 2020). While this approach is initially useful, it does not address the prediction of upcoming episodes. To predict new emerging episodes or prodromal symptoms, we need patient data before an episode and during an emerging episode with prodromal symptoms, even better also data during an episode and after. Such data would allow us to train models to discriminate between healthy, prodromal and disordered states on a within-subject level or to relate speech features to dimensional symptomatology. This would approximate the ultimate goal in clinical practice, namely to decide within a given patient that yesterday's speech features were normal, but today's speech features predict an emerging episode. Unfortunately, longitudinal studies of patients with MDD including regular speech samples, regular psychopathological ratings as ground truth and sufficient variance in this ground truth, i.e., changes in healthy and disordered states, are rare (Cummins et al., 2015; Low et al., 2020).

To address this gap, we used a longitudinal dataset in which repeated assessments of depressive momentary states and speech features derived from selfie videos were collected concomitantly by patients with an acute depressive episode (Wadle et al., 2024). While Wadle and colleagues (2024) used classical statistics (multilevel models) and focused on three specific, theory-driven speech features (speech rate, speech pauses, and pitch variability), which did indeed show associations with depression severity, we wanted to improve on several levels. Given the large number of speech features available, the aim of the present study was to extend our previous findings by examining a comprehensive set of 89 speech features and by employing more complex modeling approaches in terms of ML. We aim to contribute to this field, as we only identified three ML studies using longitudinally assessed data in a clinical (as opposed to subclinical) population with multiple data points per patient to predict depression severity based on speech features.

In one of the studies, speech samples and concomitant mood self-ratings were collected from 30 patients with MDD via AA over the course of two weeks (Gerczuk et al., 2022). ML analyses revealed a correlation of $\rho = .61$ between the actual and predicted mood scores, and an improvement in prediction when using personalized ($\rho = .79$) instead of non-personalized models.

The most promising dataset at present is from the RADAR-CNS-consortium project, with two relevant publications (Campbell et al., 2023; Cummins et al., 2023). In the paper by Cummins et al. (2023), speech data were collected in the form of a scripted task and a free-response task from 461 patients with MDD every two weeks for 18 months. A set of 28 speech features was analyzed using linear mixed models. Associations were found between elevated depression symptoms and speech rate, articulation rate, and speech intensity. However, the authors mention in their limitations that the results are based on the cohort level, which limits insights into intra-individual depression-related speech changes, which they plan to investigate in the future. The other publication from the RADAR-CNS project focused on the benefits of model personalization (Campbell et al., 2023). Data from the scripted (n = 271) and free response (n = 258) task from a subset of patients were used to explore personalized and generalized ML models. Three speech parameter sets were extracted from a total of 8,004 speech samples, with personalization proving beneficial for their binary depression classification (high vs. low depression severity). Specifically, running a support vector regression (SVR) classifier based on the extended version of the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) from the free-response task for this binary decision resulted in better performance for the personalized compared to the generalized models.

Building on previous work by the authors (Wadle et al., 2024), we aim to contribute to closing this gap and to the understanding of speech-based longitudinal monitoring of MDD. Specifically, we were interested in whether a multi-parameter ML approach would significantly improve prediction, compared to our previous study, which focused on the three most prominent speech features. In addition, we explored in which mechanism for splitting training and test data was most successful, with a particular focus on the idea of personalized prediction. To do so, we analyzed a dataset of patients (n = 30) diagnosed with MDD during sleep deprivation therapy, a fast-acting treatment (Wirz-Justice & Benedetti, 2020) that results in a significant improvement of depressive states in most of the patients within 36-hours. The given treatment ensures short-term effects, which is advantageous compared to other studies such as the RADAR—MDD project where patients are observed over two years to reveal illness episodes (Matcham et al., 2019). In Wadle et al. (2024), patients reported momentary depressive states and recorded concomitant selfie videos talking about their current feelings 2-3 times per day for up to three weeks. Speech features were extracted from the speech samples using the software open-source Speech and Music Interpretation by Large-Space Extraction (openSMILE; Eyben et al., 2010). To assess the potential clinical utility of automated symptom monitoring using speech features, we trained and evaluated a comprehensive set of > 500 ML pipelines (by optimizing hyperparameters of random chance and dummy regressors for baseline comparisons, random forest, linear regression, SVR, and eXtreme Gradient Boosting regression (XGBoost) models) to predict individual symptom severity. We used five different approaches to evaluate whether these ML models generalize across patients or whether personalized splits are superior: 1) group 5-fold cross-validation on a subject level, 2) a leave-one-subject-out (LOSO) approach, and 3) a train-test-split with 2-fold cross-validation using different splitting techniques, 3a) chronological split with the first half as training and the second half as test set, 3b) odd-even split, with chronologically sorted data put into train and test set by turns, 3c) a random split, which was repeated ten times.

Methods

Sample

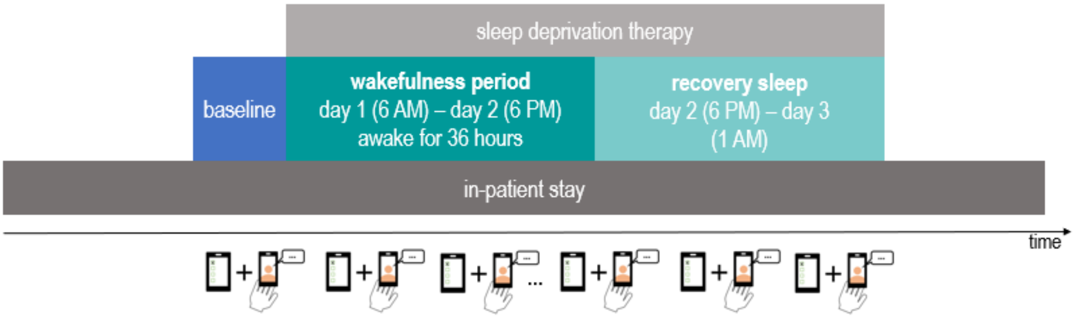
We analyzed a dataset that was collected as part of the SLEDGE II pilot study (Sleep Deprivation and Gene Expression; DRKS00022025). The initial sample consisted of 30 inpatients from the Central Institute of Mental Health in Mannheim, Germany, who experienced an acute depressive episode (ICD-10) on admission to the hospital. The final sample to be analyzed consisted of 22 patients (55% male) aged between 18 and 63 years (mean 33.5, *SD* 12.4, median 29), as the dataset of eight patients had to be excluded completely. Specifically, four patients did not record any videos; one patient did not say anything during the recordings (23 videos); the data of two patients lacked sound due to technical problems (30 videos); one patient was excluded because he or she recorded only two videos. The final sample corresponds to 18 patients with moderate depression and four patients with severe depression at study inclusion, as assessed by clinical expert interview using the Montgomery–Åsberg Depression Rating Scale (MADRS; Montgomery & Åsberg, 1979). The mean MADRS score was 28 for patients with moderate depression and 39 points for patients with severe depression. Exclusion criteria were comorbid substance use disorders and personality disorders.

Study Procedure

Data were collected by patients on a study smartphone using the movisensXS software (https://movisens.com/en). The patients underwent sleep deprivation therapy as part of their depression treatment. In other words, patients had to stay awake for approximately 36 hours. Treatment effect and relapse can be measured in a matter of four days (Wirz-Justice & Benedetti, 2020), resulting in substantial within-person variance for many patients in the dataset. After at least one day of baseline assessment, sleep deprivation therapy was conducted on what we define as day 1 (Figure 2). Specifically, patients stayed awake from 6 AM on day 1 to 6 PM on day 2. Recovery sleep was allowed from 6 PM on day 2 until 1 AM on day 3. Data were collected before, during and after sleep deprivation therapy for up to 26 days. During the first week of the study, smartphones sent prompts tree times per day (morning, afternoon, evening); in addition, self-initiated assessments were possible to report specific events or to catch up on missed assessments. To reduce patient burden, the sampling scheme was changed to two prompts per day (morning, evening). At each prompt, patients were asked to complete items about their current affective state and to record a selfie video reporting how they currently felt. Patients returned the smartphone at the end of the study.

Figure 2

Sleep Deprivation Study Design



Ambulatory Assessment: E-diary Ratings and Selfie Videos

The dataset contains three sets of momentary affect ratings in the form of e-diary ratings at each prompt. The full assessment tools are described in Wadle et al. (2024). As the analysis in the present work is limited to the target variable of momentary depression, we focus here on its detailed description. Depression severity was assessed using the short version of the Allgemeine Depressionsskala (ADS-K; Hautzinger, 1988). We adapted the ADS-K to fit the characteristics of momentary assessment with 14 items on depressive mood (excluding the sleep item) rated on a scale from 0 = *rarely* to 3 = *mostly* (Appendix A3.1). We recoded the reversed items so that higher scores indicated higher intensity of depressive symptoms, thereafter, we calculated mean values. In addition to the e-diary ratings just described, patients were asked to record selfie videos with the following instructions: *Please keep the camera stable during the recording and record your whole face. Please describe in 10 - 20 seconds how you currently feel*.

Ethical Considerations

The Ethics Committee II of the Medical Faculty Mannheim, University of Heidelberg, Germany, approved the study. Patients were informed about the aims and study procedures. All patients gave informed consent and could withdraw from the study at any time.

Data Preprocessing

Initially, the dataset contained 899 recorded selfie videos. As mentioned above, we excluded all videos of four patients (55 videos) and removed two additional videos with technical damage. We extracted audio tracks from the 842 remaining videos using the *ffmpeg* package in Python and archived them as wav files (sampling rate = 48 kHz, mono = 1 channel). In the next step, we listened to all recordings and removed test runs (*n* = 14), content-free accidental short recordings (*n* = 29), recordings in which the microphone was covered (*n* = 27), and assessments in which either the recording or the affective state assessment was missing (*n* = 24). Moreover, if two consecutive assessments occurred within 15 minutes of each other (*n* = 21), the second assessment was removed unless the audio quality of the first recording was insufficient, in which case the second assessment was kept. Finally, we excluded recordings containing third-party speech (*n* = 8) and recordings with insufficient speech intelligibility due to background noise (*n* = 9). Prior to speech parameter extraction, we filtered the remaining 710 recordings using *DeepFilterNet2* (Schröter et al., 2022) to remove background noise.

Acoustic Features

We extracted acoustic features using the functionals (v02) of *eGeMAPS* (Eyben et al., 2016) from the open-source toolkit *openSMILE* implemented in Python (Eyben et al., 2010). Given the limitless number of potential speech features and to increase comparability across studies, this minimalistic set of 88 acoustic features is recommended for use in clinical speech analysis (Eyben et al., 2016). We added the parameter speech rate, which requires transcription of the recordings. We obtained the transcript using an automatic speech recognition system according to published procedures (Abulimiti et al., 2020) and corrected the transcripts manually. To determine speech rate, we calculated the ratio of words divided by the duration of the speech sample. In our previous publication (Wadle et al., 2024), we included a subset of three of these speech features (top-down selected: F0semitoneFrom27.5Hz-_-sma3nz_stddevNorm, MeanUnvoiced-SegmentLength, words per second) in multilevel model analyses and found an association between each of them and depression severity. In the present work, however, we included all of the described 89 speech features as predictors for depression severity in our ML models.

Machine Learning

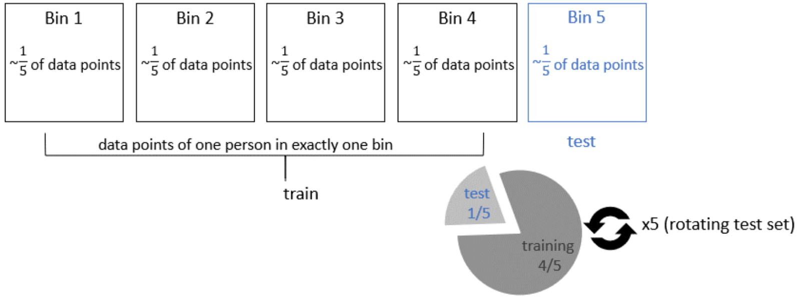
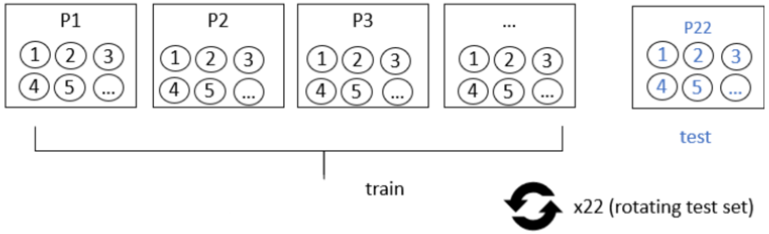
Five ML analyses were conducted to determine the optimal model for predicting ADS-K mean scores from our 89 speech features Table 6. All analyses used consistent preprocessing, including median imputation for missing data and standard scaling for feature normalization. A variety of models were evaluated: a random chance and a dummy regressor (mean and median; results of the superior are shown) for baseline comparisons, random forest, linear regression, SVR, and XGBoost regression. The models were fine-tuned using nested cross-validation and a systematic grid search to optimize the hyperparameters, ensuring the robustness and reliability of our results using the *PHOTON AI* software package (Leenings et al., 2021).

Model performance was assessed quantitatively using the R^2 score (coefficient of determination). This metric evaluates the proportion of variance in the dependent variable that can be explained by the independent variables, providing a clear measure of model effectiveness. It is essential for comparing different regression models in our analysis by quantifying how well each model explains the variability in the dataset. The performance metrics for each model and splitting technique combination were averaged to provide a comprehensive evaluation of model performance.

We also present mean absolute error (MAE) scores which measure how close the predicted and actual values are. MAEs provide a straightforward interpretation given that they are calculated in the same units as the underlying data. Clinical relevance can be inferred.

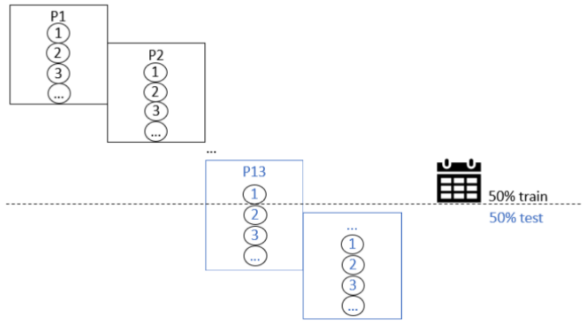
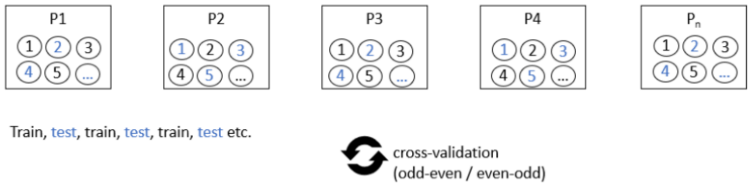
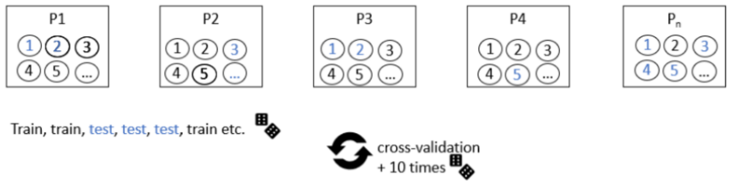
Table 6

Overview of Splitting Techniques

Split mechanism	Explanation	Visualization
Group 5-fold cross validation	Separation of data points into five bins of approximately equal size, with the condition that each patient's data are represented in exactly one bin, i.e., either in the training set or the test set, but not both. Train on all but one bin, test on the remaining bin. Repetition of the procedure until each bin has been used once as a test bin (5-fold cross-validation).	 Bin 1 ~$\frac{1}{5}$ of data points Bin 2 ~$\frac{1}{5}$ of data points Bin 3 ~$\frac{1}{5}$ of data points Bin 4 ~$\frac{1}{5}$ of data points Bin 5 ~$\frac{1}{5}$ of data points data points of one person in exactly one bin test train test 1/5 training 4/5 x5 (rotating test set)
Leave-one-subject-out	Train on data from all but one patient. Test on data from the one left-out patient. The procedure was repeated until each subject was used in the test arm (here: *n*=22).	 P1 1 2 3 4 5 ... P2 1 2 3 4 5 ... P3 1 2 3 4 5 ... ... 1 2 3 4 5 ... P22 1 2 3 4 5 ... test train x22 (rotating test set)

(continued)

Overview of Splitting Techniques (continued)

Split mechanism	Explanation	Visualization
Chronological split	Train on the chronologically first 50% of data, test on the last 50%.	 P1 1 2 3 ... P2 1 2 3 ... ... P13 1 2 3 ... ... 1 2 3 ... 50% train 50% test
Odd-even split	Odd assessment points were assigned to the training set, even assessment points to the test set. Then the implementation of a 2-fold cross-validation.	 P1 P2 P3 P4 P_n 1 2 3 4 5 ... Train, test, train, test, train, test etc. cross-validation (odd-even / even-odd)
Random split	Data points were randomly assigned to either train or test sets. This was repeated ten times with a 2-fold cross-validation calculated in each repeated run.	 P1 P2 P3 P4 P_n 1 2 3 4 5 ... Train, train, test, test, test, train etc. cross-validation + 10 times

Note. For visualizations: squares represent data bins in the first row and individual patients in the remaining rows; circles represent individual data points. P = patient.

Higher R^2 scores and lower MAE scores indicate superior model performance. *P*-values < .05 are considered to be statistically significant. Negative R^2 scores indicate poor model performance, and in such cases, the *p*-value is not of interest.

Group 5-Fold Cross-Validation

In our first analytical approach, we used group 5-fold nested cross-validation to assess model performance. Data points were divided into five bins of approximately equal size, ensuring that each patient's data appeared in only one bin, either in the training set or the test set, but not both. This means that samples from a single patient were treated as a distinct group, ensuring the integrity of individual data within each validation fold. The model was trained on four bins and tested on the remaining bin. The procedure was repeated until each bin had been used as the test bin, completing the 5-fold cross-validation. This approach tested whether the predictive patterns identified could generalize from one group of patients to another by modeling the association between speech features and depression severity across multiple patients.

Leave-one-subject out

In the second approach, we used the maximum possible data in a subject-based split for the training set. That is, we used data from all but one patient in the training set with the goal of predicting data from this one unknown patient. This reflects a potential future clinical use case where a trained model is applied to a new, unknown patient. Thus, this analysis tests whether the identified predictive pattern generalizes to an unknown patient.

In the following three approaches, we split the data fifty-fifty by using three different splitting techniques: a chronological split, an odd-even split, and a random split.

Chronological split

In this approach, we used a chronological train-test split where the first 50% of the data (355 data points), ordered by assessment date, were used as the training set and the last 50% were used as the test set (355 data points). Note that our patients were recruited over a time period of three years and two months. This means that sometimes data were collected from only one patient and sometimes from two patients at the same time. Specifically, 13 patients of our final sample were enrolled consecutively. For nine consecutive patients (i.e., nine pairs of patients), there is an overlap in assessment time when comparing the first assessment and the last assessment of an individual patient. Consequently, *earlier* patients are included in the training set, *later* patients only in the test set and three patients in both. No cross-validation was applied, as this would indicate a prediction backwards in time. This approach aimed to simulate a realistic prediction scenario by training the models on earlier assessments and testing their performance on later data points, thereby evaluating the predictive performance for future depression severity based on past assessments.

Odd-even split

This method employed a nested 2-fold cross-validation approach, in which patient-wise chronologically sorted data were alternately assigned to the training or test set based on odd and even collection points. As a result, half of the data from each patient is represented in the training set and half in the test set. Importantly, with this splitting mechanism, we assume that both the test and training sets are likely to contain data points from different states, namely severely depressed states and euthymic states right after the intervention. This approach has the advantage that the model is trained with both, individual data from depressive and euthymic states, and it avoids having all depressive data in the training set, but euthymic data only in the test set. Accordingly, this allows us to model and evaluate the predictive performance of speech features in clinical use cases. For example, predicting the severity of depression in a new depressive episode of a patient with a history of recurrent depression, who is already known by the model.

Random split

Since there is only one way to split data into training and test sets in the odd-even split, we aimed to test the replicability of these findings here. We randomly split our data into test and training sets and performed 2-fold cross-validation. There are 716 choose 358 = 1.03 × 10214 ways to randomly split the data into two halves. With this splitting mechanism, it is possible that some data points never appear in the training set. Therefore, we repeated this random split ten times and reported the mean values.

Results

Descriptive Results

Our final dataset consisted of 710 pairs of self-reported depressive momentary states and speech features extracted from concomitantly recorded selfie videos. Self-reported depression severity, as indicated by ADS-K responses (scale 0-3), was on average 1.2 (*SD* 0.6). The intraclass correlation coefficient for the ADS-K was 0.47, indicating that 53% of the variance in momentary depression symptoms is attributable to within-person variability. The reliability index of the ADS-K in the present study was excellent as evaluated according to McDonald Omega (.87 within-person and .90 between-person).

Machine Learning

We present the performance of each of our 30 ML approaches in Table 7. All combinations of our six models (from top to bottom: random chance, dummy regression, random forest regression, linear regression, SVR, XGBoost regression) and our five splitting mechanisms (from left to right: group 5-fold cross-validation, LOSO, chronological split, odd-even split, random split) are included in the table. We show R^2 scores and MAE along with their *p*-values.

Group 5-Fold Cross-Validation

In our initial analysis using group 5-fold cross-validation, all tested regressors yielded negative R^2 scores and failed to reach a performance above chance level (Table 7). This indicates that none of the models were able to significantly explain the variance of the target variable and thus failed to provide reliable predictive insights for the ADS-K mean scores in this specific setup. The models were not suitable for the dataset under the group 5-fold cross-validation scheme. This finding necessitates a reconsideration of the model parameters, feature selection, or possibly the experimental design to improve predictive performance.

Leave one subject out

The LOSO approach yielded comparable results. All models tested yielded non-significant negative R^2 scores (Table 7). This indicates that none of the models effectively explained the variance of the target variable and all models were unable to predict mean the ADS-K scores for an unknown patient in this particular setup.

Chronological split

In the chronological split analysis, none of the models achieved statistically significant results (Table 7). These results suggest that none of the models evaluated were effective in explaining the variance in the ADS-K mean scores or providing reliable predictions in this setup.

Table 7

Model Performances

Model	Splitting techniques									
	Group 5-fold cross-validation		Leave-one-subject-out		Chronological split		Odd-even-split		Random split	
	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)
Random chance	-3.306	0.920	-6.833	0.910	-2.364	0.941	-2.115	0.920	-2.205	0.890
Dummy regression (median)	-0.289 (*p* = .79)	0.499 (*p* = .41)	-3.624 (*p* = .92)	0.557 (*p* = .72)	-0.107 (*p* = .99)	0.491 (*p* = .99)	-0.001 (*p* = .35)	0.482 (*p* = .48)	-0.007 (*p* = .72)	0.488 (*p* = .84)
Random forest regression	-0.102 (*p* = .09)	0.455 (*p* = .04)	-4.392 (*p* = .99)	0.540 (*p* = .29)	-0.213 (*p* = .65)	0.519 (*p* = .81)	0.336 (*p* <.001)	0.381 (*p* <.001)	**0.305** **(*p* <.001)**	0.396 (*p* <.001)

(continued)

Model Performances (continued)

Model	Splitting techniques									
	Group 5-fold cross-validation		Leave-one-subject-out		Chronological split		Odd-even-split		Random split	
	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)	R^2 Score (*p*-value)	MAE (*p*-value)
Linear regression	-25.508 (*p* = .67)	0.588 (*p* = .50)	-37.258 (*p* = .71)	0.602 (*p* = .31)	-0.364 (*p* = .15)	0.534 (*p* = .18)	-0.179 (*p* <.001)	0.445 (*p* <.001)	-0.558 (*p* = .06)	0.459 (*p* <.001)
Support vector regression	-0.136 (*p* = .008)	0.468 (*p* = .004)	-4.006 (*p* = .88)	0.570 (*p* = .89)	-0.106 (*p* = .59)	0.439 (*p* = .87)	0.313 (*p* <.001)	0.388 (*p* <.001)	0.293 (*p* <.001)	0.401 (*p* <.001)
XGBoost regression	-0.093 (*p* = .07)	0.455 (*p* = .03)	-3.568 (*p* = .41)	0.550 (*p* = .03)	0.084 (*p* = .98)	0.442 (*p* = .99)	**0.339** **(*p* <.001)**	0.380 (*p* <.001)	0.289 (*p* <.001)	0.399 (*p* <.001)

Note. XGBoost = eXtreme gradient boosting. Superior model per splitting technique printed in bold.

Odd-even split

Overall, the performance of three models tested was above chance level (Table 7). The XGBoost regression emerged as the superior performer, achieving an R^2 score of 0.339 and an MAE of 0.38 (both $p < .001$). These results indicate that approximately 33.9% of the variance in the ADS-K mean scores can be explained by the speech features using this model. The MAE indicates that the mean difference between the predicted and the actual scores is 0.38 units on the ADS-K depression severity scale ranging from 0-3. This substantial improvement in model performance of the superior model in this approach compared to our previous ML approaches demonstrates the potential effectiveness of the XGBoost model when data are alternately assigned to training and test sets based on odd and even collection points. This analysis highlights the importance of including both depressive and euthymic data points from the same individual in both the training and test set. In addition to the XGBoost model, the SVR and random forest regression yielded statistically significant results of a descriptively comparable order of magnitude.

Random split

The random forest regression emerged as the superior performer (Table 7) in the random split. The model achieved an R^2 score of 0.305 and an MAE of 0.396 (both $p < .001$). These results indicate that using this model, approximately 30.5% of the variance in the ADS-K mean scores can be explained by the speech features. The MAE of the random forest model indicates that the mean difference between the predicted and the actual scores is 0.396 units on the ADS-K depression severity scale ranging from 0-3. In addition to the random forest regression, the SVR and XGBoost regression models reached statistical significance with descriptively comparable performance.

Discussion

The objective of this study was to evaluate if speech-based multi-parameter ML models and specific train-test splits would significantly increase the prediction of depression severity ratings compared to previous statistical analyses. Uniquely, we used a longitudinal dataset of MDD patients undergoing sleep deprivation therapy. This approach allows the observation of treatment onset and relapse within a few days, thereby allowing for a maximum of within-subject variance of momentary depressive states in our dataset. The most effective ML model (XGBoost regression with odd-even splitting) explains 33.9% of the variance of the target variable depression severity with an MAE of 0.38. It is noteworthy that this represents a 17-fold increase in predictive power over our previous analyses of this (same) dataset, which revealed an R^2_{Hox} of 2% (Wadle et al., 2024). It should be noted that in our previous analysis we focused on a subset of three speech features, whereas in the present work 89 speech features were included into the models. Furthermore, in our previous work, we used inference statistics in the form of multilevel models and ML here. The present results suggest that integrating a larger number of speech features and allowing for more complex modeling can significantly improve prediction performance. However, these findings need to be replicated in a different sample.

Moreover, our findings revealed that several models reached statistical significance, but with varying predictive power. In short, models in which both the training and the test set contained data from the same patients were successful in predicting depression severity based on speech features (odd-even split, random split). In contrast, all of our models which were tested on data from patients for whom the model was naïve, failed (chronological split, 5-fold cross-validation, LOSO). Interestingly, for both the odd-even and the random split, three ML models (random forest, SVR, XGBoost) achieved statistical significance, with an R^2 and MAE of descriptively comparable size. This suggests that these two approaches perform similarly and it is probably not critical which one is ultimately chosen. However, this conclusion must be taken with caution as we did not test the models against each other as this would require orders of magnitude more computational power than all the analyses combined here.

As noted above, all models trying to predict depression scores only of patients for whom the model was naïve, failed. This finding suggests that the predictive patterns do not appear to generalize across patients. This indicates that ML models need to be fine-tuned to the specific patient about whom predictions are to be made. This is consistent with previous research indicating better predictive performance for personalized models compared to generalized models (Campbell et al., 2023). It underscores the importance of longitudinal datasets, which are still scarce. Only when multiple data points per patient are available for training purposes, i.e., longitudinal data, can prediction reach a sufficient level.

In this context, the heterogeneity of the clinical picture of MDD must also be taken into account. Widely used diagnostic criteria allow for more than 400 possible combinations of symptoms (Goldberg, 2011; Østergaard et al., 2011). This might explain why there is no one-size-fits-all approach, i.e., associations from one patient can be easily transferred to another patient. In future work, it might be interesting to test whether models trained and tested on different patients, but with a similar clinical picture, would perform better. For example, a model trained on patients whose clinical picture is strongly characterized by having low energy might be transferable to patients with similar characteristics, but not to patients with a high degree of hyperarousal.

Limitations

Although our study demonstrates the potential use of speech features in clinical monitoring, particularly of patients with recurrent MDD, some limitations must be mentioned. First, our sample size is relatively small. However, we believe that a unique strength of our dataset is the inclusion of patients with an acute clinical diagnosis of a depressive episode requiring an inpatient stay (rather than subclinical study participants), and the true within-person design. Additionally, due to our longitudinal intervention design, we do have a relatively high number of data points per patient and a meaningful amount of variance in our target variable. Future studies are needed to test the replicability of our findings. Second, although eGeMAPS is a standardized set of speech features recommended for clinical use cases, it may not capture all relevant speech characteristics associated with depression. Nevertheless, we prefer to use predefined feature sets suggested by the community rather than creating our own features to increase the comparability across studies. In light of the previous two arguments, pooling of datasets will become very important in the future, another argument for relying on well-known feature sets. Third, we limited our analyses to five different splitting techniques, for each of which we trained over 500 ML models. Nowadays, computational power would allow us to run huge amounts of ML approaches (Winter et al., 2024). However, even with our small set of ML variants, we were still able to demonstrate the importance of individualized ML models with well-designed splitting mechanisms.

Future Directions

Although we did not test personalized ML models per se in this work, our results support the idea that personalized state-of-the-art approaches, i.e., individual ML models, are the most promising (e.g., Gerczuk et al., 2022; Wörtwein et al., 2023). A prerequisite for this is the collection of sufficient data points per person in a first step. Importantly, there must be sufficient within-person variance in illness states during this so-called burn-in phase (Kathan et al., 2022). Once a sufficient amount of data of this patient is available, a first model could be trained. As new data is coming in permanently, the model can be constantly updated with the individual's data, thus continuously improving its performance. Another idea is to start with a generalized or semi-personalized model (e.g., trained on same-sex data) to avoid the cold start problem (Kathan et al., 2022). Incoming data from the patient could be used to fine-tune the model. This is certainly a complex endeavor that requires patience and perseverance on the part of the patients, but might be worth it once a sufficiently functional model is established. In the long term, this could be particularly helpful for patients with a history of recurrent MDD. To test the feasibility of this, longitudinal studies over even longer time periods than those of the few that already exist are needed.

Moreover, to reduce patient burden, it is even more attractive to use behavioral features that patients do not have to actively collect, such as speech. Since we carry our smartphones with us most of the time anyway, and most people speak naturally in their everyday lives, these features seem promising. However, there are still many ethical and privacy questions with regard to the specific category of speech data. For example, speaker identification algorithms are needed that work reliably, on the fly, and in everyday environments (including varying background noise) to ensure that only the target's speech is analyzed.

Conclusions

Our study contributes to the emerging field of digital behavioral markers as indicators of mental health by highlighting the potential and challenges of using speech features to monitor depression. While our results suggest that speech features might be useful in predicting momentary depression severity, future research is needed to evaluate whether these findings can be replicated. Ultimately, speech-based depression monitoring systems could significantly improve patient care in the future.

CHAPTER 4

ARTICLE 3:

LINGUISTIC STYLE AS A PREDICTOR OF MOMENTARY DEPRESSION SEVERITY

This chapter is based on an adapted version of the peer-reviewed article published as

Hartnagel, L.-M., Ebner-Priemer, U. W., Foo, J. C., Streit, F., Witt, S. H., Frank, J., Limberger, M. F., Horn, A. B., Gilles, M., Rietschel, M., & Sirignano, L. (2024). Linguistic style as a digital marker for depression severity: An ambulatory assessment pilot study in patients with depressive disorder undergoing sleep deprivation therapy. *Acta Psychiatrica Scandinavica*, 1-10. https://doi.org/10.1111/acps.13726

Abstract

Background: Digital phenotyping and monitoring tools are the most promising approaches to automatically detect upcoming depressive episodes. Especially, linguistic style has been seen as a potential behavioral marker of depression, as cross-sectional studies showed, for example, less frequent use of positive emotion words, intensified use of negative emotion words, and more self-references in patients with depression compared to healthy controls. However, longitudinal studies are sparse and therefore it remains unclear whether within-person fluctuations in depression severity are associated with individuals' linguistic style.

Methods: To capture affective states and concomitant speech samples longitudinally, we used an ambulatory assessment approach sampling multiple times a day via smartphones in patients diagnosed with depressive disorder undergoing sleep deprivation therapy. This intervention promises a rapid change of affective symptoms within a short period of time, assuring sufficient variability in depressive symptoms. We extracted word categories from the transcribed speech samples using the Linguistic Inquiry and Word Count.

Results: Our analyses revealed that more pleasant affective momentary states (lower reported depression severity, lower negative affective state, higher positive affective state, (positive) valence, energetic arousal, and calmness) are mirrored in the use of less negative emotion words and more positive emotion words.

Conclusion: We conclude that a patient's linguistic style, especially the use of positive and negative emotion words, is associated with self-reported affective states and thus, is a promising feature for speech-based automated monitoring and prediction of upcoming episodes, ultimately leading to better patient care.

Introduction

Major depressive disorder (MDD) is a major health challenge and often manifests itself in a recurring or chronic condition (Vos et al., 2020). In the absence of biomarkers (Rimti et al., 2023), guiding diagnosis and treatment has traditionally relied on subjective self-report measures such as questionnaires and interviews by mental health professionals (Nezu et al., 2015). As these are administered at sporadic points in time, they are (a) prone to retrospective bias (Colombo et al., 2019) and (b) even full episodes might be missed (Ebner-Priemer & Santangelo, 2020). Patterns such as moment-to-moment fluctuations in symptoms which might be central regarding potential triggers and prodromal warning signs will remain undetected (Ebner-Priemer & Santangelo, 2020).

With the features of near real-time, continuous, active and passive data collection, the use of ambulatory assessment (AA) is advantageous as it reduces retrospective recall bias and increases ecological validity (Ebner-Priemer & Santangelo, 2020; Trull & Ebner-Priemer, 2014). AA involves the assessment of momentary self-reported symptoms and behaviors assisted by personal digital devices while patients perform their normal daily activities in their natural environment (Ebner-Priemer & Trull, 2009). One idea to reduce reliance on self-reports is to derive objective parameters from speech. Two different streams of parameters have been used in the past: acoustic and linguistic features. Multiple studies revealed differences in the acoustic dimension of speech (e.g. pitch, jitter) between depressive and non-depressive states (Wadle et al., 2024) or between depressed and healthy individuals (for review see Cummins et al., 2015; Low et al., 2020). Leveraging the development of natural language tools such as the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2003), linguistic style or the choice of words in association with MDD is also investigated. Pennebaker and colleagues (Pennebaker et al., 2015) state that individuals' everyday word use, interpreted as a behavioral manifestation of thoughts and emotions, can reveal affective and cognitive processes characteristic of mood disorders.

As the potential linguistic feature space is extensive, researchers either pursue a brute-force approach (e.g., Arevian et al., 2020; Himmelstein et al., 2018) or inform their feature selection by theoretical considerations of MDD (e.g., Huston et al., 2019; Sonnenschein et al., 2018). With regard to the latter, in many studies the use of (1) positive and negative emotion words, (2) first person singular pronouns, and (3) past tense words were analyzed. Traditional theories of depression such as Beck's Cognitive Model of Depression (Beck et al., 1979) or heightened self-focus theories (Pyszczynski & Greenberg, 1987) suggest such word use. However, empirical results are often mixed which might be due to the variety of methodological approaches and samples. Improved depressive states have been found to be associated with: (i) heightened use of positive emotion words and little use of negative emotion words (e.g., Himmelstein et al., 2018; Huston et al., 2019; Sonnenschein et al., 2018), but also greater use of both positive and negative emotion words (Weintraub et al., 2023); (ii) little use of first-person singular pronouns (review by Edwards & Holtzman, 2017; Himmelstein et al., 2018; Tackman et al., 2018), but also opposite effects (Stiles et al., 2023) and null-effects (Sonnenschein et al., 2018); (iii) little use of past tense (e.g., Habermas et al., 2008; Jones et al., 2020), but also the contrary (Weintraub et al., 2023).

In the light of future automatic everyday monitoring systems, which shall monitor the change of trajectories of patient diseases and prevent upcoming episodes, two approaches are of particular relevance: (1) longitudinal studies exploring within-person effects and (2) data collection in the field instead of controlled therapy sessions. According to current findings, only two studies meet these criteria (Arevian et al., 2020; Stiles et al., 2023). In the first study, 120 linguistic and acoustic features extracted from mental health patient recordings were analyzed (Arevian et al., 2020). Positive and negative emotion words and overall speech features were associated with global clinical assessments and depression subscores. However, only 15% of the sample were MDD patients and machine learning results refer to the full set of speech features. In another study, speech samples from MDD patients were aggregated to pre- and post-treatment assessments (Stiles et al., 2023). Decreases in negative emotion words, no difference in positive emotion words, and increases in first person pronouns for post- versus pre-treatment assessment were identified.

To inform the development of an automatic monitoring tool, robust studies in the natural environment of patients during and between state of the art assessed depressive episodes are needed. Those that exist exploit the potential of AA only weakly, for instance by aggregating multiple assessment points. Additionally, while there is a growing body of work on *inter*individual differences, studies on longitudinal *intra*individual differences during states of different depression severities are lacking. Understanding the fluctuations that occur may help to get a clearer clinical picture and provide warning of impending episodes.

With the present study, we aim to overcome the above-mentioned limitations. Based on the available evidence, the linguistic style of speech of clinically diagnosed MDD patients in association with their reported depressive state has not yet been compared longitudinally during treatment with high temporal resolution. We analyzed a dataset of speech samples and concomitant self-reported momentary affective states collected longitudinally via AA from MDD patients undergoing sleep deprivation therapy

(SDT) in an exploratory way. SDT is a chronotherapeutic intervention that can rapidly improve depression severity (Wirz-Justice & Benedetti, 2020). For our linguistic study, this is advantageous, as maximum variance in depressive symptomatology is gained within a short period of time, enabling to observe associated modifications in linguistic style within days. Finally, instead of averaging data from multiple sampling points into composite measures, we stay on the most granular level of our data collection (multiple data points per day per person), expecting to capture the dynamic ebb and flow of affective states.

To limit alpha error inflation and informed by the existing literature, we decided to include four LIWC word categories that closely resemble psychopathological phenomena. Accordingly, we formulated four hypotheses regarding associations between LIWC categories and concomitantly reported affective momentary states: With lower reported depression severity, we expected (1) more words in the LIWC category *positive emotion words*, (2) less words in the LIWC category *negative emotion words*, (3) less words in the LIWC category *first person pronouns*, and (4) less words in the LIWC category *past tense*. In addition, we analyzed associations of these categories with more broadly defined reported momentary affective states (i.e., positive and negative affective state, valence, energetic arousal, calmness). We hypothesized that the associations of LIWC categories with reported negative affective state are in the same direction as those for reported depression severity and in the opposite direction for the remaining reported affective states.

Methods

Detailed information on the sample, the study procedure and assessment tools can be found in Wadle et al. (2024) and Appendix A4.1-A4.3. Data were collected from 30 inpatients at the Central Institute of Mental Health in Mannheim, Germany, who experienced an acute depressive episode (ICD-10) at admittance to the hospital. Patients were informed about the study procedures, gave informed consent before being included in the study, and could withdraw at any time. Patients underwent SDT as part of their treatment, which involved staying awake for 36 hours. AA collection covered the period before, during, and after SDT (up to 26 days in total). In the first week of the study, 3 e-diary prompts per day (morning, afternoon, and evening) were send via smartphones, which was reduced to morning and evening prompts after the first week to minimize patient burden. E-diary prompts included a request to respond to items about the current affective state on (a) the short version of the Center for Epidemiologic Studies Depression Scale (Allgemeine Depressionsskala in German (ADS-K, Hautzinger, 1988)), (b) 15 positive and negative affective state items from an item pool (Myin-Germeys et al., 2003), and (c) a short version of the Multidimensional Mood Questionnaire (MDMQ; Wilhelm & Schoebi, 2007), and to record a selfie video to report current feelings (*Please describe in 10–20 seconds how you currently feel.*).

From originally 899 recorded selfie videos we had to exclude 155 files, mostly for technical reasons (see Appendix A4.4). We obtained transcripts of the audio tracks using an automatic speech recognition system (Abulimiti et al., 2020) and corrected them manually. The patients' transcripts were analyzed using the language processing tool LIWC (Pennebaker et al., 2015), a well-established, transparent, and reliable computer text analysis program that automatically classifies words into categories stored in a predefined dictionary (Meier et al., 2019). The output is a percentage of words allocated to the linguistic categories in relation to the total number of words in a text sample. The German LIWC 2015 version used in this study contains 18,711 words and word stems in more than 80 word categories (Meier et al., 2019). We focus on four LIWC categories: positive emotion words (*posemo*), negative emotion words (*negemo*), first person singular pronouns (*I*), and past tense (*focuspast*).

To analyze the within-subject association of LIWC categories and reported momentary affective states, we applied multilevel modeling (Snijders & Bosker, 2011) with person centered timevariant level-1 predictors (LIWC categories) and added the predictors time and time^2 in hours (centered at 2 p.m.) as covariates. We calculated separate models for each LIWC category: positive emotion words, negative emotion words, first person singular, past tense, and each affective state as outcome (reported depression severity (ADS-K), positive affective state, negative affective state, valence, energetic arousal, calmness), resulting in 24 models. As the number of speech samples was limited, we included all reported momentary affect ratings and speech recordings available irrespective of the time of assessment. We set the initial α level to 5%, applied Bonferroni corrections construct-wise (α_{adj} = .008), and performed all analyses in R (version 4.3.1 [16 June, 2023]). See Wadle et al. (2024) for details of statistical parameters and used R packages.

Results

In the final analysis, we included 744 pairs of sampled speech and affective states, an average of 34 ± 20.16 pairs per person (range = 5–87). Descriptive statistics on word categories and reported affective states, ICCs, and reliability indices (McDonald Omega and Spearman-Brown coefficients) are presented in Table 8. The upper part depicts the four LIWC categories analyzed, starting with the total word count of all speech samples with 28,128 words. Hereafter follow the four LIWC categories with the highest percentage of words found for first person singular pronouns (10.81%), followed by positive emotion words (6.22%), negative emotion words (4.64%), and past focus (3.72%). The amount of between-person variance represented by the ICCs (0.14–0.32) indicates that most of the variance is within-subject, strongly arguing for the chosen assessment and analysis approach. The lower part of Table 8 refers to the reported affective momentary states. Mean reported depression severity (ADS-K) is presented for all assessment points (1.2 ± 0.3) and for the baseline assessment only (1.5 ± 0.3), both to be classified as medium high on a scale from 0 to 3. Negative (2.4 ± 0.5) and positive (2.1 ± 0.4) affect items were medium high on average (scale 1–5), as well as MDMQ scores (valence: 44.7 ± 10.8; energetic arousal: 41.6 ± 10.6; calmness: 43.4 ± 11.45; scale 1–100). The ICCs (0.26–0.60) encourage us to have sufficient within-person variance to run multilevel regression analysis. The reliability indices were good to excellent as evaluated according to McDonald Omega and Spearman-Brown coefficients.

In Figure 3, we illustrate Pearson correlation coefficients with person-mean-centered scores. The left graph displays correlations of the six affective state variables, the right graph displays correlations of the four LIWC categories. The color scheme encodes the direction and the strength of the correlation; blue indicates a positive correlation with darker blue indicating stronger correlations; red reflects the same pattern but for negative correlations. On the y-axis and the diagonal, the names of the respective variable pairs are depicted.

Regarding correlational analyses, two main patterns emerged (Figure 3): The e-diary items on affective states showed a high coherence while the LIWC categories were weakly correlated. In detail, we found strong positive correlations (r > .5) between reported depression severity and negative affective state; strong negative correlations between reported depression severity and positive affective state, and all MDMQ

Table 8

Descriptive Statistics of LIWC Categories and Momentary Affective States

Category / variable	Output label / Scale	Examples	Total number of words	Mean (%) (*SD*; *min*[a], *max*[a])	ICC	Reliability (within)	Reliability (between)
LIWC							
Total word count	WC	-	28,128	-	-	-	-
Past focus words	Focuspast	Yesterday, said, was	1142	3.72 (3.87; 0, 7.54)	0.14	-	-
Positive emotion words	Posemo	Happy, nice, good	1609	6.22 (4.21; 0.48, 9.24)	0.17	-	-
Negative emotion words	Negemo	Sad, fear, nervous	1139	4.64 (4.87; 1.63, 14.98)	0.32	-	-
First person singular pronouns	I	I, me, mine	3009	10.81 (4.79; 7.03, 14.81)	0.21	-	-
Affective states							
ADS-K	0-3	-	-	-	0.48	0.87	0.90
ADS-K (baseline)	0-3	-	-	-	-	-	-
Negative affective	1-5	-	-	-	0.60	0.87	0.96
Positive affective	1-5	-	-	-	0.47	0.87	0.95

(continued)

Descriptive Statistics of LIWC Categories and Momentary Affective States (continued)

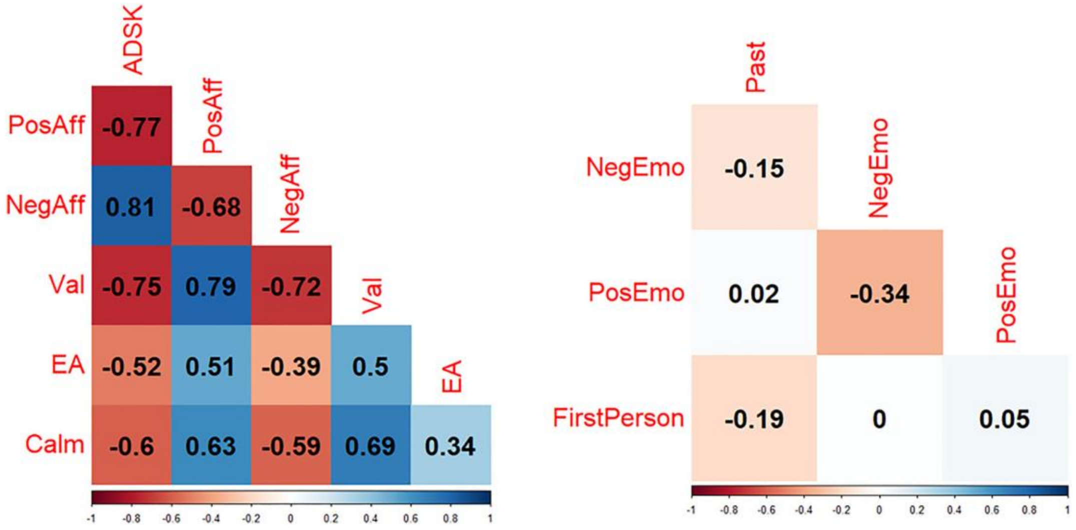
Category / variable	Output label / Scale	Examples	Total number of words	Mean (%) (*SD*; *min*[a], *max*[a])	ICC	Reliability (within)	Reliability (between)
Positive affective	1-5	-	-	-	0.47	0.87	0.95
Valence	1-100	-	-	-	0.26	0.82	0.92
Energetic arousal	1-100	-	-	-	0.29	0.61	0.84
Calmness	1-100	-	-	-	0.41	0.74	0.91

Note. [a]The mean of all patients' minimum and maximum scores. LIWC = Linguistic Inquiry and Word Count

.

Figure 3

Pearson correlations with person-mean-centered variables between affective states and between LIWC categories



Note. N between affective states and between LIWC categories (*n* between 729 and 744). Calm = calmness, EA = energetic arousal, FirstPerson = first person singular pronoun, NegAff = negative affective state, NegEmo = negative emotion words, Past = past tense; PosAff = positive affective state, PosEmo = positive emotion words, Val = valence.

subscales. The same was found for negative affective state except for a moderately strong negative correlation with energetic arousal. Correlations between positive affective state and other momentary states were opposite to the pattern found for the reported depression severity (all strong). Valence correlated strongly positively with energetic arousal and calmness. Calmness and energetic arousal correlated moderately in a positive way.

Regarding LIWC features, positive and negative emotion words showed the highest (negative) correlation, which can be classified as moderate. Past focus words correlated weakly with negative emotion words and first person singular pronouns. The other pairings were not meaningfully linked to each other. Table 9 represents the fixed effects of the separate multilevel models with positive and negative emotion words, first person singular, and past focus as statistical predictors. Momentary reported depression, positive and negative affective state are defined as outcomes. Table 10 has the same structure, but reports outcomes related to MDMQ dimensions valence, energetic arousal and calmness. All models also included the centered time and time^2 variables as statistical predictors. As findings are comparable, we report the simpler models here and expanded models in Appendix A4.5.

Positive and negative emotion words

The first two lines in Table 9 and Table 10 depict the results for the LIWC categories positive and negative emotion words as statistical predictors. Overall, there is a coherent pattern of significant associations between these two LIWC categories and momentary affective ratings. Specifically, more words in the category positive emotion words were significantly associated with less depressive symptomatology (std. β = -0.14), more positive affective state (std. β = 0.16), less negative affective state (std. β = -0.09), more (positive) valence (std. β = 0.22), more energetic arousal (std. β = 0.20), and more calmness (std. β = 0.22). The associations between negative emotion words and the momentary affective states were in the opposite direction of those just presented. Specifically, more words in the category negative emotion words were significantly associated with more severe reported depressive symptomatology (std. β = 0.16), less positive affective state (std. β = -0.18), more negative affective state (std. β = 0.15), less (positive) valence (std. β = -0.28), less energetic arousal (std. β = -0.25), and less calmness (std. β = -0.20). Effect sizes approximated with standardized beta values are comparably high with regard to the outcomes positive and negative emotion words.

Table 9

Multilevel Linear Regression Analysis to Predict Depression and Positive and Negative Affective States: Fixed Effects of LIWC Categories

Statistical predictors	E-diary ratings											
	Outcome: ADS-K				Outcome: Positive affective state				Outcome: Negative affective state			
	Beta	Stand. β	*SE*	*p*-Value	Beta	Stand. β	*SE*	*p*-Value	Beta	Stand. β	*SE*	*p*-Value
Positive emotion words	-0.02	-0.14	<0.01	**<.001**	0.03	0.16	<0.01	**<.001**	-0.02	-0.09	<0.01	**<.001**
Negative emotion words	0.02	0.16	<0.01	**<.001**	-0.03	-0.18	<0.01	**<.001**	0.03	0.15	<0.01	**<.001**
First person singular	<0.01	0.08	<0.01	.066	<-0.01	-0.04	<0.01	.214	0.01	0.07	<0.01	**.007**
Focus past	<-0.01	-0.03	<0.01	.298	<0.01	0.05	<0.01	.135	<-0.01	<-0.01	<0.01	.876

Note. ADS-K = Allgemeine Depressionsskala Kurzform. Stand. β = standardized β coefficient. Statistical significance printed in bold.

Table 10

Multilevel Linear Regression Analysis to Predict Valence, Energetic Arousal, and Calmness: Fixed Effects of LIWC Categories

Statistical predictors	E-diary ratings											
	Outcome: Valence				Outcome: Energetic arousal				Outcome: Calmness			
	Beta	Stand. β	*SE*	*p*-Value	Beta	Stand. β	*SE*	*p*-Value	Beta	Stand. β	*SE*	*p*-Value
Positive emotion words	1.12	0.22	0.18	**<.001**	0.99	0.20	0.17	**<.001**	1.19	0.22	0.17	**<.001**
Negative emotion words	-1.24	-0.28	0.17	**<.001**	-1.11	-0.25	0.15	**<.001**	-0.95	-0.20	0.16	**<.001**
First person singular	-0.21	0.05	0.17	.215	-0.06	-0.01	0.15	.713	-0.25	-0.05	0.16	.110
Focus past	0.21	0.04	0.20	.297	0.25	0.05	0.19	.181	0.42	0.07	0.19	**.028**

Note. Stand. β = standardized β coefficient. Statistical significance printed in bold.

First person singular words

Next, we show results for the LIWC category first person singular, which did not reveal such a coherent pattern as positive/ negative emotion words. First person singular was significantly associated with negative affective state (std. β = 0.07). There was a trend with respect to reported depressive symptomatology, but this result was not statistically significant (std. β = 0.08). All other associations were not significant.

Past focus words

The LIWC category past focus was not significantly associated with any of the momentary affective state variables.

Discussion

This is the first study to investigate momentary affective states and concomitant speech samples collected longitudinally via AA from patients diagnosed with MDD during SDT inducing rapid shifts in symptomatology. We found a coherent pattern for the LIWC categories positive and negative emotion words in association with concurrently reported affective states. Specifically, we found that using more positive emotion words and fewer negative ones is linked to lower reported depression severity and negative affective states. Additionally, it corresponds to higher levels of positive affective states, (positive) valence, energetic arousal, and calmness. These results suggest that changes in linguistic style extracted from ambulatorily assessed everyday speech samples may be indicative of mood changes.

As hypothesized, we found a positive association between MDD severity and the frequency of negative emotion words and a negative association between MDD severity and positive emotion words. This replicates previous work (Himmelstein et al., 2018; Huston et al., 2019) and is in line with Beck's depression model (Beck et al., 1979) suggesting that patients suffering from MDD have a severely negative attitude. Interestingly, our results generalized over the more broadly defined affective states assessed (positive and negative affective states, valence, calmness, energetic arousal). This connects to discussions on speech features being considered disorder-specific or transdiagnostic (Arevian et al., 2020; Sonnenschein et al., 2018).

With respect to our hypothesis on first person singular pronoun use, our results were more limited. An increased use was observed in association with negative affective states and a trend for depressive symptoms, but not with any of the other variables. In a previous review, a small positive correlation was found indicating higher self-focus during depressive states (Edwards & Holtzman, 2017). One possible explanation for the lack of a significant relationship between reported depression and first person singular personal pronouns lies in the nature of the task. Unlike studies that have collected language content during therapy sessions (Huston et al., 2019; Sonnenschein et al., 2018), or free spontaneous speech (Weintraub et al., 2023), patients in this study talked about their current mood in a selfie video, which might have been beneficial as more mood related content has been reported. Furthermore, earlier studies concluded that linguistic indicators of self-referencing seem to rather reflect the negative affective component of depressive symptoms. Thus, they might be rather broader indicators of negative affectivity than specifically indicating depressive states (Tackman et al., 2018).

Our fourth hypothesis, which was based on prior studies showing that MDD patients refer more frequently to the past in their narratives compared to healthy participants (Habermas et al., 2008; Stiles et al., 2023), was not supported by our data. However, in a more recent study, more past focus words were associated with less depression severity (Weintraub et al., 2023). Surprisingly and not in line with any of these previous findings, in our study the amount of past words was not associated with MDD. It is likely that our task instructions resulted in a rather low variance of past word frequencies between time points and different affective states in contrast to earlier language samples that often instructed a life review feature (Habermas et al., 2008).

Our results suggest that short speech samples collected in everyday life may serve as a behavioral marker of changes in depression severity and may to some extent compensate for the lack of objective biomarkers in future studies. The use of emotion words in particular (positive and negative) may provide clinically meaningful information that could contribute to the detection of impending episodes. Taking the advantage of the widespread availability of smartphones and smartwatches that people wear every day, speech-based monitoring of MDD is a promising approach. It should be noted that we consider such an everyday speech tool to be a supportive monitoring system, coexisting with clinical sessions which are essential. However, due to its continuous and unobtrusive applicability, speech-based monitoring could offer the crucial advantage of bridging the time between clinical sessions. Patients who are approaching their personal relapse threshold could be identified earlier, for example, if their linguistic style shows the use of more negative emotion words.

While promising in theory, the development of a speech-based monitoring system faces many challenges. We have to identify a speech sampling strategy that is (a) most informative, (b) comes with the smallest patient burden, (c) preserves the privacy of patients and bystanders, and (d) impacts smartphone battery life minimally. More specifically, it is unclear whether active sensing (i.e., asking patients to actively record a speech sample or call a system) is necessary or whether passive sensing is sufficient. Especially in the case of passive sensing, the need for privacy-preserving tools requires technological development, for example, to filter out bystander's speech or to create transcripts in real-time.

This pilot study had several limitations. First, the sample size was limited and future studies are needed to investigate replicability of results as well as generalizability beyond an inpatient sample. Still, the dataset at hand is unique as it is based on a true within-subject design with a relatively high number of data points per patient. A meaningful amount of variability in reported depression severity over a short period of time is stimulated by the study design; this is crucial from a theoretical perspective, as meaningful variance is necessary in both parameters to uncover existing associations. Second, SDT might have additional effects, such as fatigue, beyond the antidepressant one. This has to be explored in future work. Third, as we focused our analysis on four LIWC categories, it remains unclear whether other word categories are also sensitive to changing depressive states. However, as a vast variety of possible linguistic features is assess- and extractable in theory, feature selection should be made in a considered manner in order to limit alpha error inflation and to increase replicability (Wadle & Ebner-Priemer, 2023). We decided to focus on features resembling psychopathological phenomena in MDD as closely as possible and for which both a theoretical foundation and previous empirical evidence are available. Fourth, LIWC is a word count based automated language analysis and does not consider the context in which the categorized words are used or negations. However, recent findings show comparability between LIWC-based and hand-labeled (and thus *corrected*) categories (Stiles et al., 2023). LIWC categories can be compared across studies and have thus proven to be reliable and valid (Tackman et al., 2018). Fifth, although we used automated transcription, its manual correction was time-intense. For the future, especially if linguistic analysis is used in the context of AA, it is crucial to have reliable speech-to-text tools, as manual correction won't be feasible. Finally, our speech task instruction might have biased word use. Whereas this might have increased number of negative and positive words, it might have limited words related to the past. While previous LIWC studies mostly collected longer speech samples or narratives, we requested brief samples to reduce patient burden. Still identifying significant associations between linguistic style and reported depression severity supports the reliability of the measurement and is promising for AA speech sample collection, where minimizing effort per assessment is crucial for feasibility.

To conclude, our study provides evidence for associations between fluctuations in the use of positive and negative emotion words and momentary affective states. These changes happened within a relative short time, not lagging behind and as such are a real marker. We want to particularly emphasize that the sample consisted of clinically diagnosed patients with an acute depressive episode. The intervention study design involving SDT ensured a maximum of within subject dynamics of affective states within just a few days. The use of these words as a marker is promising for the development of future technology predicting upcoming episodes on an individual level. And this research adds important observations with respect to the aim of developing an automated depression monitoring system.

CHAPTER 5

ARTICLE 4:

SMART DIGITAL PHENOTYPING

This chapter is based on an adapted version of the peer-reviewed commentary published as

Wadle, L.-M., & Ebner-Priemer, U. W. (2023). Smart digital phenotyping. *European Neuropsychopharmacology*, 76, 1-2

Digital phenotyping, one of the hottest topics in psychiatry, generates an incredible amount of irrelevant and meaningless features that have little to do with clinical phenomena. Accordingly, the payoff in terms of explained variance of clinical outcomes is low. The addition of onboard real-time analytics will unlock its full potential by enabling digital phenotyping with high validity while ensuring privacy, resulting in smart digital phenotyping.

In detail, Tom Insel speculated that technology and information science will outperform neuroscience and genomics (2018), while the World Health Organization (WHO; 2019) has particularly highlighted mobile technology as a promising way to improve global mental health. All these promises refer in special to digital phenotyping (Jain et al., 2015), a method that uses data from personal digital devices to unobtrusively quantify human behavior in everyday life with high temporal resolution over long periods of time. In practice, multiple smartphone sensors can continuously track every second of a lifetime without any effort on the part of the person being tracked. This real-time data stream of app usage, typing speed, phone calls, light input and acceleration, to name just a few features, is used to predict changes in health status such as relapse or treatment response.

While the promise of digital phenotyping is immense in theory, its full potential has not been leveraged (Anmella et al., 2022) or worse, digital phenotyping parameters are often shallow and misleading (Ebner-Priemer & Santangelo, 2020). The two main reasons for this are: a) data collection tends to focus on features that are easy to collect and not on those that might be clinically meaningful (Ebner-Priemer et al., 2020); b) data collection and analysis is brute force, i.e., we collect an immense amount of data and extract an immense number of features from it, just because we can.

Let’s start with the validity of digital phenotyping features. Have you ever seen *number of unanswered incoming phones calls*, *data traffic: mobile upload in megabyte* or *number of app usage: Facebook* in our classification manuals, DSM or ICD, as a criterion for a specific disorder? Probably not. And have you ever thought about how these characteristics are clinically different from *number of unanswered outgoing phones calls*, *data traffic: upload via Wi-Fi in megabyte* or *number of app usage: TikTok*? Probably not also. So why do we measure these features? Because they are easy to measure and not because they have a strong link to our clinical phenomena. These problems are reflected in a low number of features that replicate across studies and a very low amount of explained variance (Anmella et al., 2022; Ebner-Priemer et al., 2020).

The second issue relates to the brute force manner. Extracting predefined voice features from mobile phone conversations to predict mental health states has become worthwhile. While the usual number of smartphone sensor features is a few hundred at most, the standard affect-related voice feature set is ten times larger, raising questions about alpha-error inflation. For example, using the standard 6552 features of the affect-related voice feature set (Eyben et al., 2010) in combination with classical statistical testing, could lead to 327 significant features just by chance. Using multiple outcomes, such as just a single expert interview, a single self-report, and a single dimensional expert-rating, will push the number of significant by chance findings to 1000. Although big data in mHealth research poses its own new challenges, including data management, individualized predictions and inter-subject variability, upcoming textbooks are tailored to provide guidance (Mehl et al., 2023).

It is therefore not surprising that digital phenotyping and mobile health have not yet lived up to their high expectations (Anmella et al., 2022). What we need are smart digital phenotypes. Smart digital phenotypes are both: a) defined in a smart way, by closely resembling the psychopathological phenomena themselves, and b) assessed in a smart manner. Bipolar disorder may help to illustrate this point. While the easy-to-assess features reported above (*number of unanswered incoming phone calls*, *data traffic: mobile upload in megabytes*, etc.) are somehow related to communicativeness, they do not accurately reflect the DSM-5 criterion of *more talkative than usual or feeling under pressure to talk* (American Psychiatric Association, 2013). There are smarter ways to operationalize *pressure to talk* in mobile sensing, e.g., by calculating the number of words per minute or the length of pauses between words, which are probably features we unconsciously rely on in our clinical interviews. However, recording face-to-face communication in public is challenging, raises privacy concerns, or is even prohibited by law (Fusar-Poli et al., 2022).

This calls for smart assessments: Extracting features from an incoming audio stream in real time using machine learning techniques, without storing the raw material, ensures privacy and anonymity of third parties. This has not yet been implemented in mental health, but there are promising examples from the field of affective computing. Schindler et al. (2022) have shown that social behavior and environments can be automatically classified from uncontrolled, everyday audio recordings. For example, deep learning algorithms based on the Google AudioSet transform audio-based context information into a 128-feature vector (Gemmeke et al., 2017), and in a second step this feature vector is used to classify which of the 521 possible audio classes are present in a situation. While such predicted audio classes, including crying, shouting, laughing or speech in general, are not yet classical diagnostic criteria, this work shows that the use of real-time onboard analysis mitigates privacy concerns and envisions training of more specific classifiers, such as *pressure to talk*.

Our patients deserve better than a flash-in-the-pan digital phenotyping research. The conceptual advantages of digital phenotyping are obvious: the ability to monitor symptoms at high frequency over long periods of time in order to provide early warning of relapses or upcoming episodes. We also know how to improve its effectiveness. We need to focus on clinically meaningful features and not get distracted by the immense numbers of features and findings by chance. Of course, developing meaningful clinical digital phenotypes is cumbersome, but work in affective computing shows that smart digital phenotyping can be done. This will be a meaningful step towards preventing new episodes, reducing patient burden and proving the WHO (2019) right, namely that mobile technology is a promising way to reduce global mental health burden.

CHAPTER 6

GENERAL DISCUSSION

Main Results

The overall aim of this doctoral thesis was to unravel associations between depression severity and other momentary affective states and acoustic and linguistic features extracted from everyday speech samples. For this endeavor, speech samples were collected longitudinally via AA from acutely depressed patients undergoing SDT 2-3 times per day before, during, and after therapy. This fast-acting treatment opened the opportunity to capture a maximum of variability of different affective state levels within a short period of time and to analyze them in relation to speech characteristics present at the corresponding time points. I used state-of-the-art computational tools for the extraction of acoustic and linguistic features (openSMILE and LIWC). In Article 1, using multilevel models, I revealed links between three preselected acoustic features and affective states including depression severity. Specifically, lower pitch variability, higher speech rate, and shorter speech pauses were linked to more pleasant momentary states; that is lower depression severity, higher positive affect and lower negative affect, and higher positive valence, higher energetic arousal, and higher calmness. A combined model of all three speech features resulted in R^2_{Hox} of 2%.

In Article 2, my goal was to extend the depression-related findings of Article 1 by testing whether predictive performance could be improved using multi-parameter ML methods and different train-test split scenarios. An increase in predictive performance was expected because 89 speech features were now included in the models, compared to three features in Article 1, and more complex ML modeling was used. Indeed, the superior model tested explained 33.9% of the variance in the outcome variable depression severity. This represents a 17-fold increase in predictive performance over our previous results. Testing different train-test splitting techniques revealed another finding: Models were successful in prediction only when the training and test sets contained datapoints from the same patients. Splitting techniques that lead to unknown patients in the test set, i.e., patients for which the model was naïve, did not produce significant results. Concluding, this suggests that the predictive patterns do not seem to generalize across patients, calling for fine-tuned and personalized models.

In Article 3, I shifted my focus to linguistic analysis. Analyzing the same underlying dataset, I was able to identify coherent associations between the use of positive and negative emotion word use and reported affective states. In detail, saying more positive emotion words and fewer negative emotion words was associated with more pleasant affective states in terms of lower reported depression severity and lower levels of negative affect, and higher levels of positive affect, positive valence, energetic arousal, and calmness.

In Article 4, I present a commentary advocating for smart digital phenotyping. I discuss that digital phenotyping is promising on one hand, however, until now, it has not yet reached its full potential. Often, only small (but still significant) effects are reported in studies. One explanation may be the plethora of possible features that can be measured and extracted today, and that studies often select the easily accessible low-hanging fruits rather than features closely resembling psychopathological phenomena. Second, I discuss digital phenotyping to become smarter, in terms of running real-time analytics on devices allowing for privacy preserving applications. Overall, the thoughts in this commentary round out the topic of everyday speech data in the context of mental health. An unlimited number of speech features can be extracted from speech data, which must be well thought out and handled, and moreover, the collection and analysis of speech data is an extremely challenging topic from an ethical and privacy perspective. I will take a closer look at these aspects and how they should be addressed in future research in this chapter.

The results of each of the studies presented are discussed and integrated into the current state of research in the respective articles. Viewed from a higher perspective, three main results can be derived from the synopsis of results, which will be discussed in the following.

First, the present study findings support the main assumption of this doctoral thesis that speech features, both acoustic and linguistic, are associated with depression severity. This is the first study to show this link in longitudinal high variability within-person data from currently depressed patients undergoing a fast-acting and remitting treatment (here: SDT). Overall, the acoustic features pitch variability, speech rate, and speech pauses, as well as the full set of the eGEMAPS features, and the linguistic features positive emotion words and negative emotion words were predictive of depression severity. Looking more closely at the results, one remarkable finding is the 17-fold increase in predictive power when using multi-parameter ML models based on 89 acoustic features compared to performing multilevel model regressions analysis based on three preselected acoustic features (a subset of the former). Since we changed the number of speech features and the analysis method at the same time, it is difficult to disentangle which thereof (or possibly both of them) is responsible for the immense performance gain. This unique twofold analytical approach using the same underlying dataset demonstrates the potential of ML for large numbers of variables where inference statistics reach their limits. Running multilevel model regressions with 89 speech features would result in alpha error inflation due to multiple testing, likely producing chance findings. Moreover, speech features are computationally extracted from the physical signal and abstract (e.g., MFCCS), thus not necessarily interpretable, and often difficult to grasp for researchers. As a consequence, formulating a priori hypothesis or selecting a reasonable number of features for testing can be challenging. ML opens new ways of dealing with large numbers of potential predictors which is often the case in the speech context and also to consider personalized models. It has to be kept in mind that classical ML models (and also those used in the present work) do not account for the nested data structure resulting from repeated assessments, which is a major strength of multilevel models, and cannot distinguish between within and between variance. However, recent innovative methodological developments in ML aim to do justice to the specificities of the dependent data structure (Wörtwein et al., 2023).

While the amount of variance explained may seem impressive stunning at first glance, future studies have to show whether these results can be replicated. This is particularly important because the train and test sets in our superior model contained data from the same patients and were therefore not completely independent from each other. As another finding was that the associations in our ML models were not generalizable and the benefit of personalized models was demonstrated, an ideal (though not necessarily realistic) dataset would be new data points from the same patients included in the present study. From a practical perspective in the context of future real-time ML analysis in AA, the need for sufficient computational power on the device is an important aspect to consider. This would simply not be feasible with current computing times (a single model in the present work ran for 12 hours), and the impact on battery life should also be tested. Server-based analysis may be a solution, but it would require a stable internet connection, and data security may also be an issue.

Second, because acoustic and linguistic features were analyzed separately in the present work, the question arises as to whether a combination of both types of language would be beneficial. Previous results from classification studies suggest that a combination of both can enhance performance (Morales & Levitan, 2016). With respect to our dataset, an unpublished analysis simultaneously including both sets of predictors in a multilevel regression model revealed that, with one exception, individual predictors remained significant (Hartnagel et al., upcoming conference talk). This suggests that information based on both the acoustic and linguistic dimensions each make a substantial and independent contribution, and it may be beneficial to integrate both in future analysis. From a practical point of view, linguistic analysis requires an extra step, namely the transformation of spoken words into written text. However, nowadays automatic speech-to-text tools efficiently produce transcripts. Although these transcripts may not be completely accurate, the findings by Pentland and colleagues (2023) indicate that the accuracy is sufficient to produce similar LIWC results based on manually and automatically transcribed texts. A more pressing issue may be data protection. Caution is warranted when using speech-to-text tools pre-installed on smartphones, as their terms and conditions may include sending data to the company's servers.

A third point to discuss is that acoustic and linguistic features were also linked to the other more broadly defined affective states assessed (positive and negative affect, valence, energetic arousal, calmness), not just to reported depression severity. On the one hand, this is reasonable because these underlying factors could all be part of the mosaic, whose complete picture would be labeled MDD (i.e., a person experiencing a depressive episode is likely to experience negative valence and diminished positive affect). Strong correlations (range: |.53| to .81) between ADS-K scores and these affective state variables support this notion. On the other hand, this finding calls into question the extent to which the associations are specific to MDD, or whether they should rather be interpreted as transdiagnostic features that are not specific to MDD. After all, other mental health disorders are often reflected in changes in positive and negative affect, valence, energetic arousal, and calmness are often affected as well. As the present dataset only contains data from MDD patients, it is necessary to look at studies investigating speech characteristics in other mental health disorders. Overall some features appear to be relevant across disorders, such as pitch variability in MDD, posttraumatic stress disorder, and anxiety (Low et al., 2020). Others are more specific, allowing to differentiate between compared disorders, e.g., more sadness-related words in MDD compared to anxiety disorder (Sonnenschein et al., 2018). It may be necessary to model multiple speech features rather than single ones if the goal is to capture diagnostic specificities (Arevian et al., 2020; Spruit et al., 2022).

Also potential comorbidity, which is common in MDD (Kupfer & Frank, 2003; Sonnenschein et al., 2018) has to be kept in mind. From a practical perspective, the presence of common and distinct features indicating different diagnosis, hint into the direction, that speech features are more suitable for the clinical use case described above, namely as a diagnostic aid in the process of monitoring recurrent depression. As previously mentioned, clinical visits should not be completely replaced by speech-based monitoring, because they might also be relevant for detecting comorbidity.

Limitations and Outlook

The work presented in this doctoral thesis contributes valuable steps in the long journey towards a mobile depression monitoring tool. Revealing within-person associations between speech features and depression severity was a first step in this endeavor. In this section, I provide an outlook on future research avenues to gain further insights in the context of speech-based monitoring. Partly, the limitations of the present study serve as inspiring starting points for future research ideas. To begin with, I will address feasibility and usability aspects, followed by ethical questions. Thirdly, I will expand the idea of using human speech as a proxy for depression severity to everyday audio information in general. Concluding, I will present an idea for an adaptive system that meets specific requirements for a depression monitoring tool, taking into account both ethical and practical criteria.

Feasibility perspective: Chances and obstacles

A first limiting factor in the present work is the small sample size, with usable data from only 22 of the initially 30 patients enrolled. But, the dataset at hand is unique with a total of 899 data points based on a true within-person study design. Future studies are needed to see if the results replicate and test our identified ML models. The various reasons for data exclusion can be summarized as including technical reasons and usability reasons. Note that the sampling scheme in the SDT was also changed from three assessments per day to two assessments per day based on patient feedback about high burden. Despite the fact that study participants are the most important allies of researchers, a systematic review shows that most studies do not evaluate the usability and acceptability aspects of remote monitoring tools (Simblett et al., 2018). In their study collecting speech samples in the wild, Arevian et al. (2020) also conducted an exit survey and found that 22 of 24 survey respondents described their experience as positive. Acceptance and barriers were explored more comprehensively in the recent large-scale RADAR-CNS project, which also involved remote smartphone-based speech collection (Dineley et al., 2021). A subsample of 385 individuals were invited to complete a questionnaire about their experiences during the two-year study. Major findings were that patients preferred scripted speech tasks to free speech tasks, and that depression severity significantly predicted comfort in responding. Reasons for not providing speech samples included not seeing smartphone notifications, low mood, and forgetfulness (Dineley et al., 2021). Identifying low mood as an obstacle for providing recordings is an important finding especially when the study sample consists of MDD patients. As a consequence, artificial intelligence (AI) bias can occur if training datasets consist of disproportionately more speech samples from positive than negative momentary states. Also, for a future mental health application, these low mood periods may be the most crucial ones. If no speech sample is provided, of course it cannot be analyzed, and will not be followed by any kind of action that may be necessary. However, if future studies replicate the association between low mood and response rate, this information could also be used as metadata. In such a case, providing or not providing a requested speech recording might already be informative about the current affective state, but at the same time, missing data could also be due to a variety of different reasons. Contradictory, analyses based on Ebner-Priemer et al. (2020) revealed that depression severity and upcoming depressive episodes were not significantly related to missing data entries in a sample of patients with bipolar disorder (personal communication with Ebner-Priemer; 08.08.2024). Since lack of motivation is also characteristic of MDD (Firth et al., 2017; Fleming et al., 2018), more research is needed on how to conduct studies that reduce patient burden while still collecting sufficient and useful data.

One idea in this regard is to switch from active to passive collection of speech-data. The Electronically Activated Recorder (EAR; Mehl, 2017; Mehl et al., 2001) could serve as an inspiration. The EAR is a digital audio recorder sampling ambient sounds (not restricted to speech and legal in one-party consent states in the U.S.), e.g., 30 seconds five times per hour. Its unobtrusiveness and high adherence were shown in different clinical populations (Mehl & Holleran, 2007). Passive speech collection (given informed consent of the target person and good speaker identification) is also possible and may be beneficial for patient compliance. However, research is needed to understand how such a system should be configured. There are still many open questions that may be answered differently depending on the research target. These may include but are not limited to how often and how long to sample, when and how to sample (event-, time-, interval-based, or random sampling), where to place the recording tool, and which recording tool to use.

Privacy and ethical considerations

If a speech-based depression monitoring tool were to become reality at some point, the technology per se would have potential to greatly improve depression care. However, privacy and ethical considerations are of paramount importance both in the developmental phase of such a system (e.g., data collection), as well as when algorithms are ready to be applied. First, to train and develop models, it is necessary to collect large amounts of speech data and corresponding ground truth affective states, i.e., how a person feels at that moment. Ideally, this data is ecologically valid, as aimed for in AA studies (Ebner-Priemer & Trull, 2009). In other words, it is collected under circumstances resembling the final application scenario as closely as possible. One could argue that it is a limitation of the present dataset that it was collected in a rather restricted environment as the sample consisted of inpatients being treated in a psychiatric ward. Additionally, patients were instructed to avoid recording third parties as much as possible. Although this was indeed the natural environment for these individuals at the respective time and it was not a locked ward, the radius of movement was rather small. This raises the question of the extent to which the results can be generalized to other real-world scenarios.

However, speech data collection in the wild is rather restricted in Europe. Data protection is a significant and very important topic when it comes to the collection and processing of speech data, as the voice is a valuable asset that is protected by law. Recording or processing a person’s spoken words without their consent can result in up to three years in prison under §201 of the German Criminal Code. Therefore, speech recordings are subject to strict legal protection. In Europe, this includes the General Data Protection Regulation and in Germany in addition the Telecommunications Act. They state that it is generally necessary for all parties involved to provide consent to the recording. In research projects, informed consent is usually obtained from study participants. However, especially in the context of AA, a problem may arise because it is neither possible to guarantee that no third party (without informed consent) is recorded, nor is it possible to collect informed consent from all possible bystanders. It can be speculated, that this may also be a reason why speech-based studies in real-life are still rare, at least in Germany. Limiting data collection to one-consent states in the U.S., which allow the recording as long as one party involved provides consent, or restricting data collection to controlled environments, could lead to biases in the final applications.

Technical solutions could help to enable the collection of everyday speech data also in countries with strict regulation, while ensuring privacy. One idea is to take advantage on recent developments in the speech and audio processing community. Advances in speaker identification, recognition, and separation (Radha et al., 2024;

Sharma et al., 2024; Togneri & Pullella, 2011) can help to identify the target person in a first step. The system could be trained to pass only acoustic information from the target speaker for further processing. In a second step, features of interest could be automatically extracted from the speech signal on-the-fly with tools such as openSMILE (Eyben et al., 2010) or LIWC (Pennebaker et al., 2003) adapted for mobile on-device use. Only those features are stored or further processed that do not allow a reconstruction of the speech signal itself, as they represent summary scores or frequencies.

Although research in computational speech signal processing is a fast-moving field, speaker identification in the wild poses several challenges (e.g., background noise and acoustic variability due to different microphones, environments and recording conditions) that are still subject of research and may reduce the reliability of such a system (Anidjar et al., 2024; Chauhan et al., 2024; McLaren et al., 2016; Togneri & Pullella, 2011). Moreover, in a recent study, Dumpala and colleagues (2023) showed an overlap in speech characteristics used to track MDD and those used for speaker recognition, which may affect performance. In short, there are possible solutions for everyday data collection in the wild, such as speaker identification as a first step, followed by on-device feature extraction without storing of raw audio. However, on the technical side, more research is needed for reliable and real-time speaker recognition. Additionally, more interdisciplinary research at the intersection of signal processing and psychology would be valuable to identify relevant and meaningful use cases and to configure tools in a human-centered and confidence-building way.

Although not directly linked to the limitations of the present study, another ethical aspect deserves consideration. As mentioned above, a speech-based MDD tool can be a valuable contribution in the mental health context. However, it could also become a target for misuse, such as when the AI behind it is implemented in other tools outside of the health sector (Cohen & Mello, 2019; Loch et al., 2022). Nowadays, huge amounts of data are shared publicly on social media, YouTube, and similar platforms, and voice messaging such as via WhatsApp opens up data sources to uses for which they were not intended. To oversimplify, it might be tempting for stakeholders such as insurance companies or employers to check these sources with an open-source MDD-monitoring algorithm to learn more about their candidates what those never indented to disclose (and probably do not even know themselves). Moreover, with widely used voice assistants such as Amazon's Echo, and Alexa, as well as Apple's Siri, it is often not transparent to users whether and how data is stored, further processed, or even shared with third parties, raising security and privacy concerns (Anniappa & Kim, 2021). Although users often express few privacy concerns (compared to non-users), research findings suggest that an incomplete understanding of privacy risks may play a role (e.g., Lau et al., 2018). In this regard, an important step has been taken by the European Union (EU). The EU AI act entered into force as from August 1st, 2024. It aims to set rules for AI used in the European market, with the goal of ensuring human-centered and ethical development and application of trustworthy AI, especially in sensitive areas. The regulations apply to all tools used in the EU, regardless of where they have been developed. There is reason for hope that this new regulation will deter or even prevent abuse.

Everyday audio beyond human speech

From a higher perspective, everyday audio information in general (beyond speech) can contribute to understanding and monitoring the experiences, triggers, and patterns of people living with MDD. If microphones are opened to catch speech signals anyway, acoustic scene classification (Burns et al., 2011; Ding et al., 2024) can help to learn about a person’s context without asking them. The *Google AudioSet* (Gemmeke et al., 2017) is a notable advance in this area. It is a comprehensive set of over 2.1 million 10-second YouTube sound clips. They have been manually annotated and labeled by human beings, according to a hierarchical ontology of 527 sound classes. A wide variety of sounds are represented in the ontology, reflecting the soundscapes and distinct audio events one may encounter in everyday life. These include but are not limited to different human sounds (e.g., speech, screaming, crying), acoustic environments (e.g., outdoor, inside small room, rural area), sounds of things (e.g., vehicles, tools), and different musical instruments and genres (Gemmeke et al., 2017). While the Google AudioSet represents the database, Google’s *VGGish* is a publicly available pre-trained convolutional neural network (Hershey et al., 2017). The model generates 128-dimensional features that are linked to the audio event classes but cannot be used to reconstruct the original sound they are based on. It is therefore a privacy preserving tool. In the context of MDD, it could be used to track contextual factors such as social interaction (Achterbergh et al., 2020), loneliness (Santini et al., 2015), and time spent in nature (Oh et al., 2017), all of which have been shown to play a critical role in MDD.

Everyday audio-based adaptive system

All of the above ideas and considerations can be brought together in the form of an audio-based adaptive system. As the name suggests, adaptive systems process an incoming stream of information, such as biosignal and behavioral data based on which they can adapt to the user and be continuously updated (Benke et al., 2024). The range of incoming data processed is comparable to often used parameters in AA research. These include sensor data streams (e.g., speech, physical activity), actively and passively tracked user behavior and input (e.g., app usage, phone calls, shopping history, selecting items, symptom reporting), and context information (e.g., GPS-based location information, temperature, acoustic environment). The output of adaptive systems can be manifolds, but most often it includes some kind of decision support or recommendation (Benke et al., 2024). Although the user may not be fully aware of it, adaptive systems are widely used today, especially in the consumer sector (e.g., Spotify, Amazon, booking.com). Areas of application in the health context include breaking up long periods of sitting by sending triggers to move (Giurgiu et al., 2020), supporting weight loss by providing food recommendations (Agapito et al., 2016), or selecting interventions in digitally delivered psychological treatments (Mukhiya et al., 2020). In short, adaptive systems can serve to provide tailored treatments, trigger interventions or alarms and continuously adapt thresholds thereof (Benke et al., 2024), contributing to personalized medicine as a whole. Overall, the validity of the parameters used serving as an input for adaptive systems is of paramount importance. In other words, parameters should be chosen and collected in such a way that they reflect well the phenomenon they are supposed to measure well (Benke et al., 2024; Wadle & Ebner-Priemer, 2023). Again here, human speech and everyday audio are promising candidates because they are informative about affective states, context, and human behavior (Low et al., 2020).

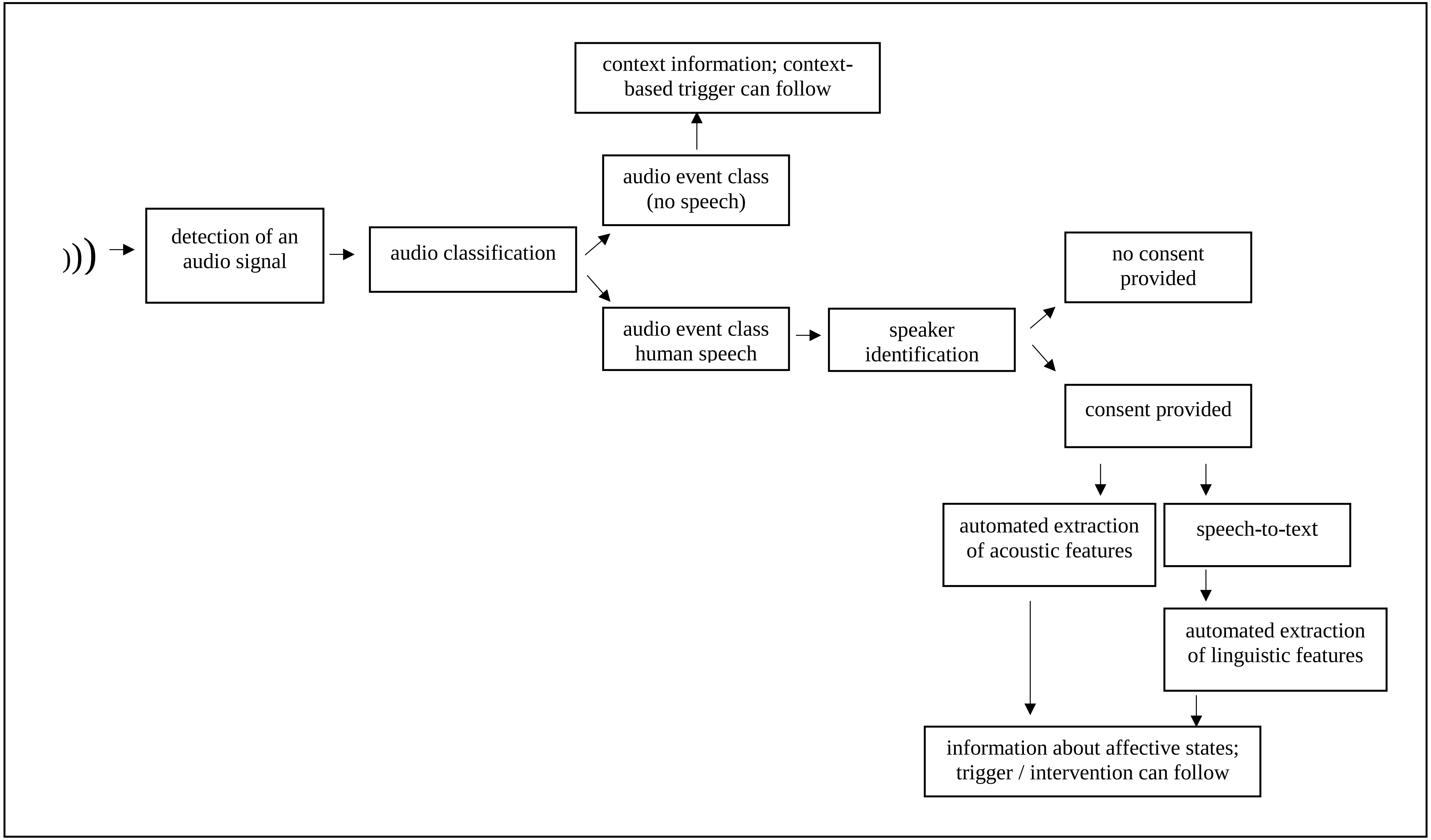
In the context of MDD, an audio-based adaptive system could be fed with information on the patient in a so-called burn in phase. The patient could provide information about stressful contexts or triggers (e.g., crowded public transportation) but also on pleasant of even curative situations (e.g., spending time in nature) that could later be identified by the system based on audio information. Additionally, speech data could be collected and ground truth about the current depression severity provided to train a personalized ML model for future identification or prediction of affective states. Like this, the system starts with some basic information which can be enriched continuously. For instance, sending short e-diary can help the system to learn more about a person's experiences and the associated feelings (e.g., that going for a run reduced symptoms). This knowledge can later on be used by the system to suggest JITAIs. Additionally, feedback loops should be implemented to keep the patient's data base as accurate as possible as preferences can change and some JITAIs may be more helpful than others. All information flows into the user's data base which is continuously updated. Trained models can also be used to predict user's future states and suggest JITAIS. For instance, if the model has learned that a person's depressive state usually deteriorates when the person is alone for a long time (which can be derived from audio data), it could suggest to call a friend.

In the following, theoretical key points of a technical processing pipeline for a fictive audio-based adaptive system will be described, illustrating its potential in the mental health sector. As mentioned earlier, ethical and privacy concerns can be obstacles for unlocking the full potential of audio-based tools, but recent advances in AI may facilitate future applications. The first step for an audio-based adaptive system is to detect the audio input and start processing. Since audio signals are pervasive, it is necessary to determine in more detail how this should be implemented. There are several possibilities, similar to classic sampling schemes in AA research: time-based sampling (e.g., every day at 8 AM), interval-based sampling (e.g., every 30 minutes for 30 seconds), random sampling (e.g., five times during the day at random times), or event-based sampling, which would require continuous passive listening until a predefined event (e.g., specific sound class) occurs. The most appropriate sampling scheme can vary greatly depending on the phenomenon of interest. Additionally, it is necessary to determine how long the microphone should be activated (i.e., the tool should *listen*). It may also be the case that no audio signal is detected because it is below a (predefined or adaptive) threshold. For instance, if a person sits at home all alone for an extended period of time. This in itself can be important information useful for an adaptive system, e.g., with the consequence that it sends triggers to animate a person for activities. Future research is necessary to determine the most promising configurations for signal detection depending on the research question or use case.

In a second step, the previously introduced audio classification comes into play (Figure 4). The purpose is twofold. On the one hand, audio event classes are identified and context information can be derived. Based on predefined contexts of interest, real-time triggering can follow in the sense of triggering specific actions or interventions or sending a short questionnaire to evaluate the situation. On the other hand, this processing step also serves as a bottleneck for further processing that is limited to speech. If the audio classifier detects human speech, it extracts speaker embeddings, which are used to identify speakers who have given informed consent or who are registered users of the application at hand. Speech samples provided during the initial use of the system could serve as a starting point for training the model to recognize target speakers. Additional persons of interest such as family members can also be registered given informed consent. These people can be identified as interaction partners (i.e., an identified social interaction could be more accurately classified as a social interaction with the mother), or speech data can also be further processed. Another possibility is to use information based on metadata, such as the share of speech time. However, the main speech data of interest is that of the target person. Given this person has been identified, the speech data can be further processed. Acoustic features can be inferred directly, and linguistic information can be extracted from a transcript via an intermediate speech-to-text processing step. Both inputs,

Figure 4

Audio Processing Pipeline


)))
detection of an audio signal
audio classification
audio event class (no speech)
context information; context-based trigger can follow
audio event class human speech
speaker identification
no consent provided
consent provided
automated extraction of acoustic features
speech-to-text
automated extraction of linguistic features
information about affective states; trigger / intervention can follow

separately or combined, can serve as indicators of affective states. Based on the results, further actions can be initiated such as sending feedback, triggering interventions such as a breathing exercise, sending alarms to therapists, or sending a short questionnaire to further explore the situation. This allows patients to gain more insight into their behavior, potential triggers, unhealthy habits, vicious cycles, but also beneficial behaviors. Overall, such adaptive systems in the healthcare context can contribute tremendously to tailored treatment.

Clearly, the walk through this fictional audio-based adaptive system is highly simplified. Quality and reliability checks have to be implemented as audio quality in the wild may be challenging depending on situation. Also sensitivity-specificity trade-offs are to be found and probably adjusted over time. Regarding speaker recognition, systems should rather be more strictly configured in order to analyze spoken data only in case a speaker is recognized with high probability. Overall, all single components of such as system have to work reliably and have to be robust to different real-life scenarios, e.g. different background noises. To sum up, there are plenty of opportunities for future researchers to explore this topic, but this theoretical vision is a first step in this endeavor. Interdisciplinary projects would be extremely helpful, as a comprehensive understanding of all aspects may only be possible by integrating different communities, such as ethics, audio signaling, psychology, psychiatry, and usability experts.

Conclusion

To sum up, with this doctoral thesis I have contributed to the understanding of the association between everyday speech characteristics and momentary affective states, in particular depression severity. The presented analyses are based on a longitudinal dataset of patients with an acute depressive episode that are treated with SDT. Using AA, speech samples and concomitant momentary affective states were collected multiple times before, during, and after the intervention. I used multilevel modeling and ML to reveal associations between extracted acoustic and linguistic features and reported momentary states. It can be concluded, that speech-based information can serve as a diagnostic aid in the long-term monitoring of MDD.

REFERENCES

Abd-Alrazaq, A., AlSaad, R., Aziz, S., Ahmed, A., Denecke, K., Househ, M., Farooq, F., & Sheikh, J. (2023). Wearable Artificial Intelligence for Anxiety and Depression: Scoping Review. *Journal of Medical Internet Research*, *25*, e42672. https://doi.org/10.2196/42672

Abulimiti, A., Weiner, J., & Schultz, T. (2020). Automatic Speech Recognition for ILSE-Interviews: Longitudinal Conversational Speech Recordings Covering Aging and Cognitive Decline. *Interspeech 2020*, 3795–3799. https://doi.org/10.21437/Interspeech.2020-2829

Achterbergh, L., Pitman, A., Birken, M., Pearce, E., Sno, H., & Johnson, S. (2020). The experience of loneliness among young people with depression: A qualitative meta-synthesis of the literature. *BMC Psychiatry*, *20*(1), 415. https://doi.org/10.1186/s12888-020-02818-3

Agapito, G., Calabrese, B., Guzzi, P. H., Cannataro, M., Simeoni, M., Care, I., Lamprinoudi, T., Fuiano, G., & Pujia, A. (2016). DIETOS: A recommender system for adaptive diet monitoring and personalized food suggestion. *2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 1–8. https://doi.org/10.1109/WiMOB.2016.7763190

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Anidjar, O. H., Marbel, R., & Yozevitch, R. (2024). Harnessing the power of Wav2Vec2 and CNNs for Robust Speaker Identification on the VoxCeleb and LibriSpeech

Datasets. *Expert Systems with Applications*, *255*, 124671. https://doi.org/10.1016/j.eswa.2024.124671

Anmella, G., Faurholt-Jepsen, M., Hidalgo-Mazzei, D., Radua, J., Passos, I. C., Kapczinski, F., Minuzzi, L., Alda, M., Meier, S., Hajek, T., Ballester, P., Birmaher, B., Hafeman, D., Goldstein, T., Brietzke, E., Duffy, A., Haarman, B., López-Jaramillo, C., Yatham, L. N., … Kessing, L. V. (2022). Smartphone-based interventions in bipolar disorder: Systematic review and meta-analyses of efficacy. A position paper from the International Society for Bipolar Disorders (ISBD) Big Data Task Force. *Bipolar Disorders*, *24*(6), 580–614. https://doi.org/10.1111/bdi.13243

Anniappa, D., & Kim, Y. (2021). Security and Privacy Issues with Virtual Private Voice Assistants. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 0702–0708. https://doi.org/10.1109/CCWC51732.2021.9375964

Arevian, A. C., Bone, D., Malandrakis, N., Martinez, V. R., Wells, K. B., Miklowitz, D. J., & Narayanan, S. (2020). Clinical state tracking in serious mental illness through computational analysis of speech. *PLOS ONE*, *15*(1), e0225695. https://doi.org/10.1371/journal.pone.0225695

Beck, A. T., Rush, J., Shaw, B. E., & Emery, G. (1979). *Cognitive Therapy of Depression* (n). The Guilford Press.

Benasi, G., Fava, G. A., & Guidi, J. (2021). Prodromal Symptoms in Depression: A Systematic Review. *Psychotherapy and Psychosomatics*, *90*(6), 365–372. https://doi.org/10.1159/000517953

Benke, I., Knierim, M., Adam, M., Beigl, M., Dorner, V., Ebner-Priemer, U., Herrmann, M., Klarmann, M., Maedche, A., Nafziger, J., Nieken, P., Pfeiffer, J., Puppe, C., Putze, F., Scheibehenne, B., Schultz, T., & Weinhardt, C. (2024). Hybrid Adaptive Systems. *Business & Information Systems Engineering*, *66*(2), 233–247. https://doi.org/10.1007/s12599-024-00861-y

Ben-Zeev, D., & Young, M. A. (2010). Accuracy of Hospitalized Depressed Patients' and Healthy Controls' Retrospective Symptom Reports: An Experience Sampling Study. *Journal of Nervous & Mental Disease*, *198*(4), 280–285. https://doi.org/10.1097/NMD.0b013e3181d6141f

Boland, E. M., Rao, H., Dinges, D. F., Smith, R. V., Goel, N., Detre, J. A., Basner, M., Sheline, Y. I., Thase, M. E., & Gehrman, P. R. (2017). Meta-analysis of the antidepressant effects of acute sleep deprivation. *The Journal of Clinical Psychiatry*, *78*(8), 893.

Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., & Mohr, D. C. (2011). Harnessing Context Sensing to Develop a Mobile Intervention for Depression. *Journal of Medical Internet Research*.

Caligiuri, M. P., & Ellwanger, J. (2000). Motor and cognitive aspects of motor retardation in depression. *Journal of Affective Disorders*, *57*(1–3), 83–93. https://doi.org/10.1016/S0165-0327(99)00068-3

Campbell, E. L., Dineley, J., Conde, P., Matcham, F., White, K. M., Oetzmann, C., Simblett, S., Bruce, S., Folarin, A. A., & Wykes, T. (2023). Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models. *INTERSPEECH 2023*, *2023*, 1738–1742. https://discovery.ucl.ac.uk/id/eprint/10176456/

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., & Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and Cognition*, *56*(1), 30–35. https://doi.org/10.1016/j.bandc.2004.05.003

Chauhan, N., Isshiki, T., & Li, D. (2024). Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies. *Acoustics*, *6*(2), 439–469. https://doi.org/10.3390/acoustics6020024

Cohen, I. G., & Mello, M. M. (2019). Big data, big tech, and protecting patient privacy. *JAMA*, *322*(12), 1141–1142. https://doi.org/10.1001/jama.2019.11365

Colombo, D., Fernández-Álvarez, J., Patané, A., Semonella, M., Kwiatkowska, M., García-Palacios, A., Cipresso, P., Riva, G., & Botella, C. (2019). Current State and Future Directions of Technology-Based Ecological Momentary Assessment and Intervention for Major Depressive Disorder: A Systematic Review. *Journal of Clinical Medicine*, *8*(4), 465. https://doi.org/10.3390/jcm8040465

Csikszentmihalyi, M., & Larson, R. (1987). Validity and Reliability of the Experience-Sampling Method. *Journal of Nervous and Mental Disease*, *175*, 526–537. https://doi.org/10.1007/978-94-017-9088-8_3

Cummins, N., Dineley, J., Conde, P., Matcham, F., Siddi, S., Lamers, F., Carr, E., Lavelle, G., Leightley, D., White, K. M., Oetzmann, C., Campbell, E. L., Simblett, S., Bruce, S., Haro, J. M., Penninx, B. W. J. H., Ranjan, Y., Rashid, Z., Stewart, C., … Hotopf, M. (2023). Multilingual markers of depression in remotely collected speech samples: A preliminary analysis. *Journal of Affective Disorders*, *341*, 128–136. https://doi.org/10.1016/j.jad.2023.08.097

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, *71*, 10–49. https://doi.org/10.1016/j.specom.2015.03.004

Dallaspezia, S., & Benedetti, F. (2015). Sleep Deprivation Therapy for Depression. In P. Meerlo, R. M. Benca, & T. Abel (Eds.), *Sleep, Neuronal Plasticity and Brain Function* (pp. 483–502). Springer. https://doi.org/10.1007/7854_2014_363

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., Mohr, D. C., Dobson, R., & Hotopf, M. (2022). Digital health tools for the passive monitoring of depression: A systematic review of methods. *Npj Digital Medicine*, *5*(1), 3. https://doi.org/10.1038/s41746-021-00548-8

Demet, E. M., Chicz-Demet, A., Fallon, J. H., & Sokolski, K. N. (1999). Sleep deprivation therapy in depressive illness and parkinson's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *23*(5), 753–784. https://doi.org/10.1016/S0278-5846(99)00039-1

Dineley, J., Lavelle, G., Leightley, D., Matcham, F., Siddi, S., Peñarrubia-María, M. T., White, K. M., Ivan, A., Oetzmann, C., Simblett, S., Dawe-Lane, E., Bruce, S., Stahl, D., Ranjan, Y., Rashid, Z., Conde, P., Folarin, A. A., Haro, J. M., Wykes, T., … The RADAR-CNS Consortium, -. (2021). Remote Smartphone-Based Speech Collection: Acceptance and Barriers in Individuals with Major Depressive Disorder. *Interspeech 2021*, 631–635. https://doi.org/10.21437/Interspeech.2021-1240

Ding, B., Zhang, T., Wang, C., Liu, G., Liang, J., Hu, R., Wu, Y., & Guo, D. (2024). Acoustic scene classification: A comprehensive survey. *Expert Systems with Applications*, *238*, 121902. https://doi.org/10.1016/j.eswa.2023.121902

Dumpala, S. H., Dikaios, K., Rodriguez, S., Langley, R., Rempel, S., Uher, R., & Oore, S. (2023). Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. *Scientific Reports*, *13*(1), 11155. https://www.nature.com/articles/s41598-023-35184-7

Eaton, W. W., Neufeld, K., Chen, L.-S., & Cai, G. (2000). A Comparison of Self-report and Clinical Diagnostic Interviews for Depression: Diagnostic Interview Schedule and Schedules for Clinical Assessment in Neuropsychiatry in the Baltimore Epidemiologic Catchment Area Follow-up. *Archives of General Psychiatry*, *57*(3), 217. https://doi.org/10.1001/archpsyc.57.3.217

Ebner-Priemer, U., & Santangelo, P. (2020). Digital phenotyping: Hype or hope? *The Lancet Psychiatry*, *7*(4), 297–299. https://doi.org/10.1016/S2215-0366(19)30380-3

Ebner-Priemer, U. W., Mühlbauer, E., Neubauer, A. B., Hill, H., Beier, F., Santangelo, P. S., Ritter, P., Kleindienst, N., Bauer, M., Schmiedek, F., & Severus, E. (2020). Digital phenotyping: Towards replicable findings with comprehensive assessments and integrative models in bipolar disorders. *International Journal of Bipolar Disorders*, *8*(1), 35. https://doi.org/10.1186/s40345-020-00210-4

Ebner-Priemer, U. W., & Trull, T. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075

Ebrahimi, O. V., Burger, J., Hoffart, A., & Johnson, S. U. (2021). Within- and across-day patterns of interplay between depressive symptoms and related psychopathological processes: A dynamic network approach during the COVID-19 pandemic. *BMC Medicine*, *19*(1), 317. https://doi.org/10.1186/s12916-021-02179-y

Echizenya, M., Suda, H., Takeshima, M., Inomata, Y., & Shimizu, T. (2013). Total sleep deprivation followed by sleep phase advance and bright light therapy in drugresistant mood disorders. *Journal of Affective Disorders*, *144*(1–2), 28–33. https://doi.org/10.1016/j.jad.2012.06.022

Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, *68*, 63–68.

Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*(4), 637–642. https://doi.org/10.1007/s00038-012-0416-3

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462. https://doi.org/10.1145/1873951.1874246

Fahrenberg, J., & Myrtek, M. (Eds.). (1996). *Ambulatory assessment: Computer-assisted psychological and psychophysiological methods in monitoring and field studies*. Hogrefe & Huber Publishers.

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory Assessment—Monitoring Behavior in Daily Life Settings. *European Journal of Psychological Assessment*, *23*(4), 206–213. https://doi.org/10.1027/1015-5759.23.4.206

Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017). The efficacy of smartphone-based mental health interventions for depressive symptoms: A meta-analysis of randomized controlled trials. *World Psychiatry*, *16*(3), 287–298. https://doi.org/10.1002/wps.20472

Fleming, T., Bavin, L., Lucassen, M., Stasiak, K., Hopkins, S., & Merry, S. (2018). Beyond the Trial: Systematic Review of Real-World Uptake and Engagement With Digital Self-Help Interventions for Depression, Low Mood, or Anxiety. *Journal of Medical Internet Research*, *20*(6), e199. https://doi.org/10.2196/jmir.9275

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, *47*(7), 829–837. https://doi.org/10.1109/10.846676

Fried, E. I. (2017). Moving forward: How depression heterogeneity hinders progress in treatment and research. *Expert Review of Neurotherapeutics*, *17*(5), 423–425. https://doi.org/10.1080/14737175.2017.1307737

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, *1*(6), 358–368. https://doi.org/10.1038/s44159-022-00050-2

Fried, E. I., & Robinaugh, D. J. (2020). Systems all the way down: Embracing complexity in mental health research. *BMC Medicine*, *18*(1), 205, s12916-020-01668-w. https://doi.org/10.1186/s12916-020-01668-w

Fusar-Poli, P., Manchia, M., Koutsouleris, N., Leslie, D., Woopen, C., Calkins, M. E., Dunn, M., Tourneau, C. L., Mannikko, M., Mollema, T., Oliver, D., Rietschel, M., Reininghaus, E. Z., Squassina, A., Valmaggia, L., Kessing, L. V., Vieta, E., Correll, C. U., Arango, C., & Andreassen, O. A. (2022). Ethical considerations for precision psychiatry: A roadmap for research and clinical practice. *European*

Neuropsychopharmacology, *63*, 17–34. https://doi.org/10.1016/j.euroneuro.2022.08.001

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio Set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

Gerczuk, M., Triantafyllopoulos, A., Amiriparian, S., Kathan, A., Bauer, J., Berking, M., & Schuller, B. (2022). Personalised deep learning for monitoring depressed mood from speech. *2022 E-Health and Bioengineering Conference (EHB)*, 1–5. https://ieeexplore.ieee.org/abstract/document/9991737/

Giedke, H., & Schwärzler, F. (2002). Therapeutic use of sleep deprivation in depression. *Sleep Medicine Reviews*, *6*(5), 361–377. https://doi.org/10.1053/smrv.2002.0235

Giurgiu, M., Koch, E. D., Plotnikoff, R. C., Ebner-Priemer, U. W., & Reichert, M. (2020). Breaking Up Sedentary Behavior Optimally to Enhance Mood. *Medicine & Science in Sports & Exercise*, *52*(2), 457–465. https://doi.org/10.1249/MSS.0000000000002132

Global Burden of Disease Collaborative Network. (2020). *Global Burden of Disease Study 2019 (GBD 2019) results.* Institute for Health Metrics and Evaluation.

Goldberg, D. (2011). The heterogeneity of “major depression.” *World Psychiatry*, *10*(3), 226–228. https://doi.org/10.1002/j.2051-5545.2011.tb00061.x

Greenberg, P., Chitnis, A., Louie, D., Suthoff, E., Chen, S.-Y., Maitland, J., Gagnon-Sanschagrin, P., Fournier, A.-A., & Kessler, R. C. (2023). The Economic Burden of Adults with Major Depressive Disorder in the United States (2019). *Advances in Therapy*, *40*(10), 4460–4479. https://doi.org/10.1007/s12325-023-02622-x

Habermas, T., Ott, L.-M., Schubert, M., Schneider, B., & Pate, A. (2008). Stuck in the past: Negative bias, explanatory style, temporal order, and evaluative perspectives in life narratives of clinically depressed individuals. *Depression and Anxiety*, *25*(11), E121–E132. https://doi.org/10.1002/da.20389

Hare, D. L., Toukhsati, S. R., Johansson, P., & Jaarsma, T. (2014). Depression and cardiovascular disease: A clinical review. *European Heart Journal*, *35*(21), 1365–1372. https://doi.org/10.1093/eurheartj/eht462

Hartnagel, L.-M., Ebner-Priemer, U. W., Foo, J. C., Streit, F., Witt, S. H., Frank, J., Limberger, M. F., Horn, A. B., Gilles, M., Rietschel, M., & Sirignano, L. (upcoming). Linguistic style as a digital marker for depression severity: An ambulatory assessment pilot study in patients with depressive disorder undergoing sleep deprivation therapy. *DGPS*.

Hashim, N. W., Wilkes, M., Salomon, R., Meggs, J., & France, D. J. (2017). Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores. *Journal of Voice*, *31*(2), 256.e1-256.e6. https://doi.org/10.1016/j.jvoice.2016.06.006

Hautzinger, M. (1988). Ein Depressionsmessinstrument für Untersuchungen in der Allgemeinbevölkerung [A depression measurement instrument for assessing the general population]. *Diagnostika*, *34*, 167–173.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., &

Wilson, K. (2017). *CNN Architectures for Large-Scale Audio Classification* (arXiv:1609.09430). arXiv. http://arxiv.org/abs/1609.09430

Himmelstein, P., Barb, S., Finlayson, M. A., & Young, K. D. (2018). Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. *PLOS ONE*, *13*(11), e0207814. https://doi.org/10.1371/journal.pone.0207814

Horwitz, A. G., Zhao, Z., & Sen, S. (2023). Peak-end bias in retrospective recall of depressive symptoms on the PHQ-9. *Psychological Assessment*, *35*(4), 378–381. https://doi.org/10.1037/pas0001219

Horwitz, R., Quatieri, T. F., Helfer, B. S., Yu, B., Williamson, J. R., & Mundt, J. (2013). On the relative importance of vocal source, system, and prosody in human depression. *2013 IEEE International Conference on Body Sensor Networks*, 1–6. https://doi.org/10.1109/BSN.2013.6575522

Hox, J., & van de Schoot, R. (2013). Robust methods for multilevel analysis. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 387–402). Sage Publishers.

Huston, J., Meier, S., Faith, M., & Reynolds, A. (2019). Exploratory study of automated linguistic analysis for progress monitoring and outcome assessment. *Counselling and Psychotherapy Research*, *19*(3), 321–328. https://doi.org/10.1002/capr.12219

Insel, T. R. (2018). Digital phenotyping: A global tool for psychiatry. *World Psychiatry*, *17*(3), 276–277. https://doi.org/10.1002/wps.20550

Jablonka, E., Ginsburg, S., & Dor, D. (2012). The co-evolution of language and emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2152–2159. https://doi.org/10.1098/rstb.2012.0117

Jain, S. H., Powers, B. W., Hawkins, J. B., & Brownstein, J. S. (2015). The digital phenotype. *Nature Biotechnology*, *33*(5), 462–463. https://doi.org/10.1038/nbt.3223

Jones, L. S., Anderson, E., Loades, M., Barnes, R., & Crawley, E. (2020). Can linguistic analysis be used to identify whether adolescents with a chronic illness are depressed? *Clinical Psychology & Psychotherapy*, *27*(2), 179–192. https://doi.org/10.1002/cpp.2417

Jorm, A. F., Patten, S. B., Brugha, T. S., & Mojtabai, R. (2017). Has increased provision of treatment reduced the prevalence of common mental disorders? Review of the evidence from four countries. *World Psychiatry*, *16*(1), 90–99. https://doi.org/10.1002/wps.20388

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science*, *4*(6), 401–405. https://doi.org/10.1111/j.1467-9280.1993.tb00589.x

Kappas, A., Hess, U., & Scherer, K. R. (1991). Voice and emotion. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 200–238). Cambridge University Press.

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*(12), 1174–1179. https://doi.org/10.1038/mp.2012.105

Kathan, A., Harrer, M., Küster, L., Triantafyllopoulos, A., He, X., Milling, M., Gerczuk, M., Yan, T., Rajamani, S. T., Heber, E., Grossmann, I., Ebert, D. D., & Schuller, B. W. (2022). Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Frontiers in Digital Health*, *4*, 964582. https://doi.org/10.3389/fdgth.2022.964582

Kihlstrom, J. F., Eich, E., Sandbrand, D., & Tobias, B. A. (2000). Emotion and memory: Implications for self-report. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report* (pp. 81–99). Taylor & Francis.

Kohn, R., Saxena, S., Levav, I., & Saraceno, B. (2004). The treatment gap in mental health care. *Bulletin of the World Health Organization*, *82*(11), 858–866.

Koops, S., Brederoo, S. G., De Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2023). Speech as a Biomarker for Depression. *CNS & Neurological Disorders - Drug Targets*, *22*(2), 152–160. https://doi.org/10.2174/1871527320666211213125847

Kraepelin, E. (1921). *Manic-depressive insanity and paranoia*. E. & S. Livingstone.

Kupfer, D. J., & Frank, E. (2003). Comorbidity in depression. *Acta Psychiatrica Scandinavica*, *108*(s418), 57–60. https://doi.org/10.1034/j.1600-0447.108.s418.12.x

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. *New Directions for Methodology of Social and Behavioral Science*, *15*(15), 41–56.

Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–31. https://doi.org/10.1145/3274371

Leenings, R., Winter, N. R., Plagwitz, L., Holstein, V., Ernsting, J., Sarink, K., Fisch, L., Steenweg, J., Kleine-Vennekate, L., Gebker, J., Emden, D., Grotegerd, D., Opel, N., Risse, B., Jiang, X., Dannlowski, U., & Hahn, T. (2021). PHOTONAI—A Python API for rapid machine learning model development. *PLOS ONE*, *16*(7), e0254062. https://doi.org/10.1371/journal.pone.0254062

Leibenluft, E., & Wehr, T. A. (1992). Is sleep deprivation useful in the treatment of depression? *The American Journal of Psychiatry*, *149*(2), 159–168.

Lenderking, W. R., Hu, M., Tennen, H., Cappelleri, J. C., Petrie, C. D., & Rush, A. J. (2008). Daily process methodology for measuring earlier antidepressant response. *Contemporary Clinical Trials*, *29*(6), 867–877. https://doi.org/10.1016/j.cct.2008.05.012

Lerner, D., Adler, D. A., Chang, H., Lapitsky, L., Hood, M. Y., Perissinotto, C., Reed, J., McLaughlin, T. J., Berndt, E. R., & Rogers, W. H. (2004). Unemployment, Job Retention, and Productivity Loss Among Employees With Depression. *Psychiatric Services*, *55*(12), 1371–1378. https://doi.org/10.1176/appi.ps.55.12.1371

Loch, A. A., Lopes-Rocha, A. C., Ara, A., Gondim, J. M., Cecchi, G. A., Corcoran, C. M., Mota, N. B., & Argolo, F. C. (2022). Ethical Implications of the Use of Language Analysis Technologies for the Diagnosis and Prediction of Psychiatric Disorders. *JMIR Mental Health*, *9*(11), e41014. https://doi.org/10.2196/41014

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116. https://doi.org/10.1002/lio2.354

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-1881.1.3.86

Malhi, G. S., & Mann, J. J. (2018). Depression. *The Lancet*, *392*(10161), 2299–2312. https://doi.org/10.1016/S0140-6736(18)31948-2

Matcham, F., Barattieri Di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., Bwjh, P., Ranjan, Y., Rashid, Z., Rintala, A., … on behalf of the RADAR-CNS consortium. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry*, *19*(1), 72. https://doi.org/10.1186/s12888-019-2049-z

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., De Girolamo, G., Difrancesco, S., Haro, J. M., Horsfall, M., Ivan, A., Lavelle, G., Li, Q., Lombardini, F., Mohr, D. C., Narayan, V. A., Oetzmann, C., Penninx, B. W. J. H., Bruce, S., … on behalf of the RADAR-CNS consortium. (2022). Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC Psychiatry*, *22*(1), 136. https://doi.org/10.1186/s12888-022-03753-1

Mathers, C. D., & Loncar, D. (2006). Projections of Global Mortality and Burden of Disease from 2002 to 2030. *PLoS Medicine*, *3*(11), e442. https://doi.org/10.1371/journal.pmed.0030442

McLaren, M., Ferrer, L., Castan, D., & Lawson, A. (2016). The 2016 Speakers in the Wild Speaker Recognition Evaluation. *Interspeech 2016*, 823–827. https://doi.org/10.21437/Interspeech.2016-1137

Mehl, M. R. (2006). The lay assessment of subclinical depression in daily life. *Psychological Assessment*, *18*(3), 340–345. https://doi.org/10.1037/1040-3590.18.3.340

Mehl, M. R. (2017). The Electronically Activated Recorder (EAR): A Method for the Naturalistic Observation of Daily Social Behavior. *Current Directions in Psychological Science*, *26*(2), 184–190. https://doi.org/10.1177/0963721416680611

Mehl, M. R., Eid, M., Wrzus, C., Harari, G. M., & Ebner-Priemer, U. W. (2023). *Mobile sensing in psychology: Methods and applications*. The Guilford Press.

Mehl, M. R., & Holleran, S. E. (2007). An empirical analysis of the obtrusiveness of and participants' compliance with the electronically activated recorder (EAR). *European Journal of Psychological Assessment*, *23*(4), 248–257.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517–523. https://doi.org/10.3758/BF03195410

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). *"LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE- LIWC2015*. OSF. https://doi.org/10.31234/osf.io/uq8zt

Miloyan, B., & Fried, E. (2017). A reassessment of the relationship between depression and all-cause mortality in 3,604,005 participants from 293 studies. *World Psychiatry*, *16*(2), 219–220. https://doi.org/10.1002/wps.20439

Moffitt, T. E., Caspi, A., Taylor, A., Kokaua, J., Milne, B. J., Polanczyk, G., & Poulton, R. (2010). How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective *versus* retrospective ascertainment. *Psychological Medicine*, *40*(6), 899–909. https://doi.org/10.1017/S0033291709991036

Montgomery, S. A., & Åsberg, M. (1979). A New Depression Scale Designed to be Sensitive to Change. *British Journal of Psychiatry*, *134*(4), 382–389. https://doi.org/10.1192/bjp.134.4.382

Morales, M. R., & Levitan, R. (2016). Speech vs. text: A comparative analysis of features for depression detection systems. *2016 IEEE Spoken Language*

Technology Workshop (SLT), 136–143. https://doi.org/10.1109/SLT.2016.7846256

Mukhiya, S. K., Wake, J. D., Inal, Y., & Lamo, Y. (2020). Adaptive Systems for Internet-Delivered Psychological Treatments. *IEEE Access*, *8*, 112220–112236. https://doi.org/10.1109/ACCESS.2020.3002793

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics*, *20*(1), 50–64. https://doi.org/10.1016/j.jneuroling.2006.04.001

Mundt, J. C., Vogel, A. P., Feltner, D. E., & Lenderking, W. R. (2012). Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biological Psychiatry*, *72*(7), 580–587. https://doi.org/10.1016/j.biopsych.2012.03.015

Myin-Germeys, I., Peeters, F., Havermans, R., Nicolson, N. A., DeVries, M. W., Delespaul, P., & Van Os, J. (2003). Emotional reactivity to daily life stress in psychosis and affective disorder: An experience sampling study. *Acta Psychiatrica Scandinavica*, *107*(2), 124–131. https://doi.org/10.1034/j.1600-0447.2003.02025.x

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine*, *52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

Nezu, A. M., McClure, K. S., & Nezu, C. M. (2015). The Assessment of Depression. In I. H. Gotlib & C. L. Hammen (Eds.), *Treating Depression* (2nd ed., pp. 44–68). The Guilford Press. https://doi.org/10.1002/9781119114482.ch2

Oh, B., Lee, K. J., Zaslawski, C., Yeung, A., Rosenthal, D., Larkey, L., & Back, M. (2017). Health and well-being benefits of spending time in forests: Systematic review. *Environmental Health and Preventive Medicine*, *22*(1), 71. https://doi.org/10.1186/s12199-017-0677-9

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, *41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Østergaard, S. D., Jensen, S. O. W., & Bech, P. (2011). The heterogeneity of the depressive syndrome: When numbers get serious. *Acta Psychiatrica Scandinavica*, *124*(6), 495–496. https://doi.org/10.1111/j.1600-0447.2011.01744.x

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pentland, S. J., Fuller, C. M., Spitzley, L. A., & Twitchell, D. P. (2023). Does accuracy matter? Methodological considerations when using automated speech-to-text for social science research. *International Journal of Social Research Methodology*, *26*(6), 661–677.

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression.

Psychological Bulletin, *102*(1), 122–138. https://doi.org/10.1037/0033-2909.102.1.122

Quatieri, T. F., & Malyska, N. (2012). Vocal-source biomarkers for depression: A link to psychomotor activity. *Interspeech 2012*, 1059–1062. https://doi.org/10.21437/Interspeech.2012-311

Radha, K., Bansal, M., & Pachori, R. B. (2024). Speech and speaker recognition using raw waveform modeling for adult and children's speech: A comprehensive review. *Engineering Applications of Artificial Intelligence*, *131*, 107661. https://doi.org/10.1016/j.engappai.2023.107661

Reichert, M., Gan, G., Renz, M., Braun, U., Brüßler, S., Timm, I., Ma, R., Berhe, O., Benedyk, A., Moldavski, A., Schweiger, J. I., Hennig, O., Zidda, F., Heim, C., Banaschewski, T., Tost, H., Ebner-Priemer, U. W., & Meyer-Lindenberg, A. (2021). Ambulatory assessment for precision psychiatry: Foundations, current developments and future avenues. *Experimental Neurology*, *345*, 113807. https://doi.org/10.1016/j.expneurol.2021.113807

Rimti, F. H., Shahbaz, R., Bhatt, K., & Xiang, A. (2023). A review of new insights into existing major depressive disorder biomarkers. *Heliyon*, *9*(8), e18909. https://doi.org/10.1016/j.heliyon.2023.e18909

Santini, Z. I., Koyanagi, A., Tyrovolas, S., Mason, C., & Haro, J. M. (2015). The association between social relationships and depression: A systematic review. *Journal of Affective Disorders*, *175*, 53–65. https://doi.org/10.1016/j.jad.2014.12.049

Santomauro, D. F., Mantilla Herrera, A. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., Bang-Jensen, B. L., Bertolacci, G. J., Bloom, S. S., Castellano, R., Castro, E., Chakrabarti, S., Chattopadhyay, J., Cogen, R. M., Collins, J. K., … Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312), 1700–1712. https://doi.org/10.1016/S0140-6736(21)02143-7

Santor, D. A., Gregus, M., & Welch, A. (2006). FOCUS ARTICLE: Eight Decades of Measurement in Depression. *Measurement: Interdisciplinary Research & Perspective*, *4*(3), 135–155. https://doi.org/10.1207/s15366359mea0403_1

Saragoussi, D., Christensen, M. C., Hammer-Helmich, L., Rive, B., Touya, M., & Haro, J. M. (2018). Long-term follow-up on health-related quality of life in major depressive disorder: A 2-year European cohort study. *Neuropsychiatric Disease and Treatment*, *Volume 14*, 1339–1350. https://doi.org/10.2147/NDT.S159276

Schindler, D., Spors, S., Demiray, B., & Krüger, F. (2022). Automatic Behavior Assessment from Uncontrolled Everyday Audio Recordings by Deep Learning. *Sensors*, *22*(22), 8617. https://doi.org/10.3390/s22228617

Schröter, H., Maier, A., Escalante-B, A. N., & Rosenkranz, T. (2022). Deepfilternet2: Towards Real-Time Speech Enhancement on Embedded Devices for Full-Band Audio. *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 1–5. https://doi.org/10.1109/IWAENC53105.2022.9914782

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., & Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Interspeech 2013*, 148–152. https://doi.org/10.21437/Interspeech.2013-56

Scull, A. (2021). American psychiatry in the new millennium: A critical appraisal. *Psychological Medicine*, *51*(16), 2762–2770. https://doi.org/10.1017/S0033291721001975

Sharma, R., Govind, D., Mishra, J., Dubey, A. K., Deepak, K. T., & Prasanna, S. R. M. (2024). Milestones in speaker recognition. *Artificial Intelligence Review*, *57*(3), 58. https://doi.org/10.1007/s10462-023-10688-w

Simblett, S., Greer, B., Matcham, F., Curtis, H., Polhemus, A., Ferrão, J., Gamble, P., & Wykes, T. (2018). Barriers to and Facilitators of Engagement With Remote Measurement Technology for Managing Health: Systematic Review and Content Analysis of Findings. *Journal of Medical Internet Research*, *20*(7), e10480. https://doi.org/10.2196/10480

Smith, M., Dietrich, B. J., Bai, E., & Bockholt, H. J. (2020). Vocal pattern detection of depression among older adults. *International Journal of Mental Health Nursing*, *29*(3), 440–449. https://doi.org/10.1111/inm.12678

Snijders, T. A. B., & Bosker, R., J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). Sage Publishers.

Sobocki, P., Jönsson, B., Angst, J., & Rehnberg, C. (2006). Cost of depression in Europe. *The Journal of Mental Health Policy and Economics*, *9*(2), 87–98.

Solhan, M. B., Trull, T. J., Jahng, S., & Wood, P. K. (2009). Clinical assessment of affective instability: Comparing EMA indices, questionnaire reports, and retrospective recall. *Psychological Assessment*, *21*(3), 425–436. https://doi.org/10.1037/a0016869

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., & Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive Behaviour Therapy, 47*(4), 315–327. https://doi.org/10.1080/16506073.2017.1419505

Spruit, M., Verkleij, S., De Schepper, K., & Scheepers, F. (2022). Exploring Language Markers of Mental Health in Psychiatric Stories. *Applied Sciences, 12*(4), 2179. https://doi.org/10.3390/app12042179

Stiles, L., Frazier, A., & Eddington, K. M. (2023). What were you Thinking? A Comparison of Rater Coding and word Counts for Content Analysis of Thought Samples in Depression. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*. https://doi.org/10.1007/s10942-023-00507-0

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, *16*(3), 199–202. https://doi.org/10.1093/abm/16.3.199

Stone, A., Shiffman, S., Atienza, A., & Nebeling, L. (2007). *The science of real-time data capture: Self-reports in health research*. Oxford University Press.

Streit, F., Zillich, L., Frank, J., Kleineidam, L., Wagner, M., Baune, B. T., Klinger-König, J., Grabe, H. J., Pabst, A., Riedel-Heller, S. G., Schmiedek, F., Schmidt, B., Erhardt, A., Deckert, J., NAKO Investigators, Rietschel, M., Berger, K., Lieb, W., Becher, H., … Hoffmann, W. (2023). Lifetime and current depression in the German National Cohort (NAKO). *The World Journal of Biological Psychiatry*, *24*(10), 865–880. https://doi.org/10.1080/15622975.2021.2014152

Sundberg, J. (1998). Expressivity in singing. A review of some recent investigations. *Logopedics Phoniatrics Vocology, 23*(3), 121–127.

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., Edwards, T. S., Pennebaker, J. W., & Mehl, M. R. (2018). Depression, Negative Emotionality, and Self-Referential Language: A Multi-Lab, Multi-

Measure, and Multi-Language-Task Research Synthesis. *Journal of Personality and Social Psychology: Personality Processes and Individual Differences*, *116*(5), 817–834.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Thibaut, F. (2018). Controversies in psychiatry. *Dialogues in Clinical Neuroscience, 20*(3), 151–152. https://doi.org/10.31887/DCNS.2018.20.3/fthibaut

Togneri, R., & Pullella, D. (2011). An Overview of Speaker Identification: Accuracy and Robustness Issues. *IEEE Circuits and Systems Magazine*, *11*(2), 23–61. https://doi.org/10.1109/MCAS.2011.941079

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health*, *3*(2), e16. https://doi.org/10.2196/mental.5165

Torous, J., Onnela, J.-P., & Keshavan, M. (2017). New dimensions and new tools to realize the potential of RDoC: Digital phenotyping via smartphones and connected devices. *Translational Psychiatry*, *7*(3), e1053–e1053. https://doi.org/10.1038/tp.2017.25

Torous, J., Staples, P., & Onnela, J.-P. (2015). Realizing the Potential of Mobile Mental Health: New Methods for New Data in Psychiatry. *Current Psychiatry Reports*, *17*(8), 61. https://doi.org/10.1007/s11920-015-0602-0

Trevino, A. C., Quatieri, T. F., & Malyska, N. (2011). Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 42. https://doi.org/10.1186/1687-6180-2011-42

Trifu, R. N., Nemeș, B., Bodea-Hațegan, C., & Cozman, D. (2017). Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *Journal of Evidence-Based Psychotherapies*, *17*(1), 105–128. https://doi.org/10.24193/jebp.2017.1.7

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory Assessment. *Annual Review of Clinical Psychology*, *9*(1), 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

Trull, T. J., & Ebner-Priemer, U. (2014). The Role of Ambulatory Assessment in Psychological Science. *Current Directions in Psychological Science*, *23*(6), 466–470. https://doi.org/10.1177/0963721414550706

Verduijn, J., Verhoeven, J. E., Milaneschi, Y., Schoevers, R. A., Van Hemert, A. M., Beekman, A. T. F., & Penninx, B. W. J. H. (2017). Reconsidering the prognosis of major depressive disorder across diagnostic boundaries: Full recovery is the exception rather than the rule. *BMC Medicine*, *15*(1), 215. https://doi.org/10.1186/s12916-017-0972-8

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abolhassani, H., Aboyans, V., Abrams, E. M., Abreu, L. G., Abrigo, M. R. M., Abu-Raddad, L. J., Abushouk, A. I., … Murray, C. J. L. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9

Wadle, L.-M., & Ebner-Priemer, U. W. (2023). Smart digital phenotyping. *European Neuropsychopharmacology*, *76*, 1–2. https://doi.org/10.1016/j.euroneuro.2023.07.002

Wadle, L.-M., Ebner-Priemer, U. W., Foo, J. C., Yamamoto, Y., Streit, F., Witt, S. H., Frank, J., Zillich, L., Limberger, M. F., Ablimit, A., Schultz, T., Gilles, M., Rietschel, M., & Sirignano, L. (2024). Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study. *JMIR Mental Health*, *11*, e49222. https://doi.org/10.2196/49222

Weintraub, M. J., Posta, F., Ichinose, M. C., Arevian, A. C., & Miklowitz, D. J. (2023). Word usage in spontaneous speech as a predictor of depressive symptoms among youth at high risk for mood disorders. *Journal of Affective Disorders*, *323*, 675–678. https://doi.org/10.1016/j.jad.2022.12.047

Wells, J. E., & Horwood, L. J. (2004). How accurate is recall of key symptoms of depression? A comparison of recall and longitudinal reports. *Psychological Medicine*, *34*(6), 1001–1011. https://doi.org/10.1017/S0033291703001843

Wildgruber, D., Ackermann, H., & Grodd, W. (2001). Differential Contributions of Motor Cortex, Basal Ganglia, and Cerebellum to Speech Motor Control: Effects of Syllable Repetition Rate Evaluated by fMRI. *NeuroImage*, *13*(1), 101–109. https://doi.org/10.1006/nimg.2000.0672

Wilhelm, P., & Schoebi, D. (2007). Assessing Mood in Daily Life. *European Journal of Psychological Assessment*, *23*(4), 258–267. https://doi.org/10.1027/1015-5759.23.4.258

Williams, S. Z., Chung, G. S., & Muennig, P. A. (2017). Undiagnosed depression: A community diagnosis. *SSM - Population Health*, *3*, 633–638. https://doi.org/10.1016/j.ssmph.2017.07.012

Winter, N. R., Blanke, J., Leenings, R., Ernsting, J., Fisch, L., Sarink, K., Barkhau, C., Emden, D., Thiel, K., Flinkenflügel, K., Winter, A., Goltermann, J., Meinert, S., Dohm, K., Repple, J., Gruber, M., Leehr, E. J., Opel, N., Grotegerd, D., … Hahn, T. (2024). A Systematic Evaluation of Machine Learning–Based Biomarkers for Major Depressive Disorder. *JAMA Psychiatry*, *81*(4), 386. https://doi.org/10.1001/jamapsychiatry.2023.5083

Wirz-Justice, A., & Benedetti, F. (2020). Perspectives in affective disorders: Clocks and sleep. *European Journal of Neuroscience*, *51*(1), 346–365. https://doi.org/10.1111/ejn.14362

Wirz-Justice, A., & Van Den Hoofdakker, R. H. (1999). Sleep deprivation in depression: What do we know, where do we go? *Biological Psychiatry*, *46*(4), 445–453. https://doi.org/10.1016/S0006-3223(99)00125-0

World Health Organization. (2019). *WHO guideline: Recommendations on digital interventions for health system strengthening Web Supplement 2: Summary of findings and GRADE tables*. World Health Organization, Geneva.

Wörtwein, T., Allen, N. B., Sheeber, L. B., Auerbach, R. P., Cohn, J. F., & Morency, L.-P. (2023). Neural Mixed Effects for Nonlinear Personalized Predictions. *International Conference on Multimodal Interaction*, 445–454. https://doi.org/10.1145/3577190.3614115

Yang, Y., Fairbairn, C., & Cohn, J. F. (2013). Detecting Depression Severity from Vocal Prosody. *IEEE Transactions on Affective Computing*, *4*(2), 142–150. https://doi.org/10.1109/T-AFFC.2012.38

Zarate, D., Stavropoulos, V., Ball, M., De Sena Collier, G., & Jacobson, N. C. (2022). Exploring the digital footprint of depression: A PRISMA systematic literature review of the empirical evidence. *BMC Psychiatry*, *22*(1), 421. https://doi.org/10.1186/s12888-022-04013-y

Zupan, Z., Žeželj, I., & Andjelković, I. (2017). Memory Bias in Depression: Effects of Self-Reference and Age. *Journal of Social and Clinical Psychology*, *36*(4), 300–315. https://doi.org/10.1521/jscp.2017.36.4.300

APPENDIX

Supplementary Materials Chapter 2

A2.1 List of medications

Patients received guideline-compliant pharmacotherapy for depression. Five participants received monotherapy with an antidepressant agent (n=1 sertraline (SSRI); n=2 venlafaxine (SSNRI); n=1 bupropion and n=1 trazodone); thirteen patients were treated with two antidepressants (n=4 venlafaxine and trazodone; n=4 bupropion and trazodone; n=2 venlafaxine and mirtazapine; n=1 bupropione and mirtazapine; n=1 sertraline and trazodone; n=1 venlafaxine and amitriptyline; two patients were prescribed an antidepressant plus augmentation therapy (n=1 bupropion and quetiapine; n=1 venlafaxine and pregablin); one patient received a quindruple combination of two antidepressants medication (bupropion and venlafaxine) and augmentation with quetiapine, lamotrigine and pregabalin. Above-mentioned sedative, respectively, sleep-inducing medication (trazodone, amitriptyline, quetiapine, pregabalin,) was paused before sleep deprivation night.

Sertraline = SSRI

Venlafaxine = SSNRI

Mirtazapine = NaSSA

Bupropion = SNDRI

Agomelatine = MASSA

Amitriptyline = TCA

Trazodone = chemically different antidepressant

A2.2 ADS-K items with English translations in italics

1) Während der letzten Minuten haben mich Dinge beunruhigt, die mir sonst nichts ausmachen.
During the last few minutes, things that normally don't bother me worried me.

2) Während der letzten Minuten konnte ich meine trübsinnige Laune nicht loswerden, obwohl mich meine Freunde / Familie / Mitpatienten versuchten aufzumuntern.
During the last few minutes, I couldn't get rid of my gloomy mood, although my friends / family / fellow patients tried to cheer me up.

3) Während der letzten Minuten hatte ich Mühe mich zu konzentrieren.
During the last few minutes I had trouble concentrating.

4) Während der letzten Minuten war ich deprimiert / niedergeschlagen.
During the last few minutes I was depressed / down.

5) Während der letzten Minuten war alles anstrengend für mich.
During the last minutes everything was exhausting for me.

6) Während der letzten Minuten dachte ich, mein Leben ist ein einziger Fehlschlag.
During the last minutes I thought my life was one big failure.

7) Während der letzten Minuten hatte ich Angst.
During the last minutes I was afraid.

8) Während der letzten Minuten war ich fröhlich gestimmt.
During the last minutes I was in a cheerful mood.

9) Während der letzten Minuten habe ich weniger als sonst geredet.
During the last minutes I talked less than usual.

10) Während der letzten Minuten fühlte ich mich einsam.
During the last minutes I felt lonely.

11) Während der letzten Minuten habe ich das Leben genossen.
During the last minutes I enjoyed life.

12) Während der letzten Minuten war ich traurig.
During the last minutes I felt sad.

13) Während der letzten Minuten hatte ich das Gefühl, dass mich die Leute nicht leiden können.
During the last minutes I felt that people didn't like me.

14) Während der letzten Minuten konnte ich mich zu nichts aufraffen.
During the last minutes I couldn't get myself up to do anything.

A2.3 MDMQ items with English translations in italics

Im Moment fühle ich mich …	*At the moment I feel ...*
unzufrieden – zufrieden	*discontent- content*
unwohl – wohl	*unwell - well*
müde – wach	*tired - awake*
Im Moment fühle ich mich …	*At the moment I feel ...*
energielos – energiegeladen	*without energy - full of energy*
unruhig – ruhig	*agitated - calm*
angespannt – entspannt	*tense - relaxed*

A2.4 Positive and negative affect items with English translations in italics

Im Moment fühle ich mich

fröhlich / zufrieden / tatkräftig / enthusiastisch / entspannt / glücklich; einsam / traurig / unsicher; ängstlich / niedergeschlagen / schuldig / deprimiert / misstrauisch / gereizt

At the moment I feel

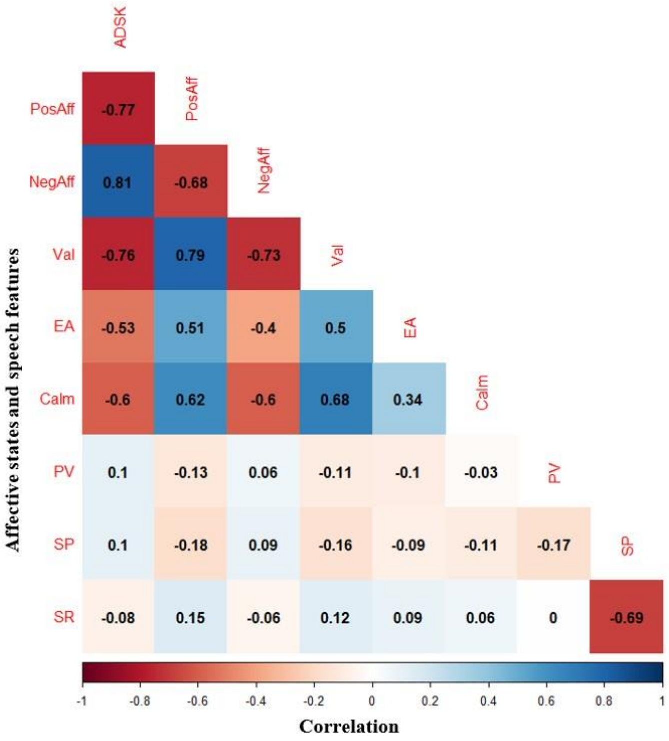
cheerful / content / energetic / enthusiastic / relaxed / happy;

lonely, sad, insecure, anxious, depressed, low-spirited, guilty, distrustful, irritable

A2.5 Additional eGeMAPS features included in exploratory analysis

spectralFlux_sma3_amean, spectralFlux_sma3_stddevNorm, spectralFluxUV_sma3nz_amean, spectralFluxV_sma3nz_amean, spectralFluxV_sma3nz_stddevNorm, mfcc1_sma3_amean, mfcc1_sma3_stddevNorm, mfcc2_sma3_amean, mfcc2_sma3_stddevNorm, mfcc3_sma3_amean, mfcc3_sma3_stddevNorm, mfcc4_sma3_amean, mfcc4_sma3_stddevNorm, mfcc1V_sma3nz_amean, mfcc1V_sma3nz_stddevNorm, mfcc2V_sma3nz_amean, mfcc2V_sma3nz_stddevNorm, mfcc3V_sma3nz_amean, mfcc3V_sma3nz_stddevNorm, mfcc4V_sma3nz_amean, mfcc4V_sma3nz_stddevNorm, equivalentSoundLevel_dBp, F2bandwidth_sma3nz_amean¸ F2bandwidth_sma3nz_stddevNorm, F3bandwidth_sma3nz_amean, F3bandwidth_sma3nz_stddevNorm

A2.6 Pearson correlations between and within affective scores and speech features



Note. n between 698 and 716. PosAff: positive affect; NegAff: negative affect, Val: valence; EA: energetic arousal; Calm: calmness; PV: pitch variability; SP: speech pauses; SR: speech rate.

A2.7 Multilevel linear regression analysis

Table 11

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects for Pitch Variability, Time, and Time^2

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.27	-	0.10	12.87	<.001
Time	<0.01	-	<0.01	0.40	.69
Time^2	<0.01	-	<0.01	-0.09	.93
Pitch variability	0.88	0.14	0.32	2.73	.007
Positive affective state					
Intercept	2.10	-	0.13	16.78	<.001
Time	<-0.01	-	<0.01	-0.97	.33
Time^2	< 0.01	-	< 0.01	-0.13	.90
Pitch variability	-1.50	-0.18	0.42	-3.56	< .001
Negative affective state					
Intercept	2.45	-	0.16	14.86	<.001
Time	<0.01	0.04	<0.01	1.74	.08
Time^2	<0.01	-	<0.01	-1.43	.15
Pitch variability	0.85	0.08	0.43	1.95	.052
Valence					
Intercept	43.72	-	2.70	16.21	<.001
Time	<0.01	-	<0.01	1.23	.22
Time^2	<0.01	<0.01	<0.01	1.67	.098
Pitch variability	-36.50	-0.16	13.61	-2.68	<.008
Energetic arousal					
Intercept	42.82	-	2.71	15.79	<.001
Time	<-0.01	0.11	<0.01	-3.46	<.001
Time^2	<0.01	<0.01	<0.01	-4.41	<.001
Pitch variability	-33.21	-0.15	12.48	-2.66	<.001
Calmness					
Intercept	40.97	-	.3.39	12.08	<.001
Time	<0.01	-	<0.01	0.20	.84
Time^2	<0.01	<0.01	<0.01	2.49	.01
Pitch variability	-11.52	-0.05	12.82	-.90	.37

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Table 12

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects for Speech Pauses, Time, and Time^2

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.27	-	0.10	12.99	<.001
Time	<0.01	-	<0.01	0.26	.79
Time^2	<0.01	-	<0.01	-0.48	.63
Speech pauses	0.52	0.10	0.18	2.80	.005
Positive affective state					
Intercept	2.09	-	0.13	16.64	<.001
Time	<-0.01	-	<0.01	-0.90	.37
Time^2	<0.01	-	<0.01	0.61	.54
Speech pauses	-1.16	-0.17	0.24	-4.84	<.001
Negative affective state					
Intercept	2.46	-	0.16	14.95	<.001
Time	<0.01	0.04	<0.01	1.75	.08
Time^2	<0.01	<0.01	<0.01	-1.90	.06
Speech pauses	0.76	0.09	0.25	3.05	.002
Valence					
Intercept	43.26	-	2.69	16.06	<.001
Time	<0.01	-	<0.01	1.28	.20
Time^2	<0.01	<0.01	<0.01	2.22	.03
Speech pauses	-34.06	-0.19	7.71	-4.42	<.001
Energetic arousal					
Intercept	42.71	-	2.71	15.74	<.001
Time	<-0.01	0.11	<0.01	-3.25	.001
Time^2	<0.01	<0.01	<0.01	-4.17	<.001
Speech pauses	-14.06	-0.08	7.14	-1.97	.049
Calmness					
Intercept	40.58		3.39	11.98	<.001
Time	<0.01	-	<0.01	0.06	.95
Time^2	<0.01	<0.01	<0.01	2.98	.003
Speech pauses	-24.27	-0.12	7.27	-3.34	<.001

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Table 13

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects for Speech Rate, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.27	-	0.10	12.93	<.001
Time	<0.01	-	<0.01	0.26	.80
Time²	<0.01	-	<0.01	-0.22	.83
Speech rate	-0.11	-0.10	0.05	-2.27	.02
Positive affective state					
Intercept	2.10	-	0.13	16.71	<.001
Time	<-0.01	-	<0.01	-0.91	.362
Time²	<0.01	-	<0.01	0.18	.859
Speech rate	0.26	0.18	0.06	4.09	<.001
Negative affective state					
Intercept	2.45	-	0.16	14.88	<.001
Time	<0.01	0.04	<0.01	1.69	.09
Time²	<0.01	-	<0.01	-1.57	.12
Speech rate	-0.13	-0.08	0.07	-2.05	.04
Valence					
Intercept	43.56	-	2.70	16.13	<.001
Time	<0.01	-	<0.01	1.28	.20
Time²	<0.01	<0.01	<0.01	1.85	.07
Speech rate	6.49	0.17	2.03	3.20	.001
Energetic arousal					
Intercept	42.77	-	2.71	15.76	<.001
Time	<-0.01	0.11	<0.01	-3.32	<.001
Time²	<0.01	<0.01	<0.01	-4.29	<.001
Speech rate	4.13	0.11	1.87	2.22	.027
Calmness					
Intercept	40.86	-	3.39	12.05	<.001
Time	<0.01	-	<0.01	0.14	.89
Time²	<0.01	<0.01	<0.01	2.63	.009
Speech rate	3.43	0.09	1.91	1.80	.07

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

A2.8 Multilevel linear regression analysis (exploratory)

Table 14

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects for Equivalent Sound Level, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.29	-	0.10	13.11	<.001
Time	<0.01	-	<0.01	0.53	.60
Time²	<0.01	-	<0.01	-1.06	.29
Equivalent sound level	-0.03	1.52	<0.01	-5.83	<.001
Positive affective state					
Intercept	2.08	-	0.13	16.59	<.001
Time	<-0.01	-	<0.01	-1.06	.29
Time²	<0.01	-	<0.01	0.96	.34
Equivalent sound level	0.05	-1.91	<0.01	6.53	<.001
Negative affective state					
Intercept	2.47	-	0.16	15.02	<.001
Time	<0.01	-	<0.01	1.92	.06
Time²	<0.01	-	<0.01	-2.27	.02
Equivalent sound level	-0.04	1.28	<0.01	-4.81	<.001
Valence					
Intercept	43.12	-	2.70	16.01	<.001
Time	<0.01	-	<0.01	1.32	.19
Time²	<0.01	-	<0.01	2.38	.02
Equivalent sound level	1.09	-1.54	0.23	4.63	<.001
Energetic arousal					
Intercept	42.31	-	2.71	15.64	<.001
Time	<-0.01	-	<0.01	-3.43	<.001
Time²	<0.01	-	<0.01	-3.66	<.001
Equivalent sound level	0.95	-1.37	0.22	4.43	<.001
Calmness					
Intercept	40.48	-	3.39	11.95	<.001
Time	<0.01	-	<0.01	0.09	.93
Time²	<0.01	-	<0.01	3.09	.002
Equivalent sound level	0.76	-1.04	0.22	3.50	<.001

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Table 15

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects for Spectral Flux, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.28	-	0.10	13.02	<.001
Time	<0.01	-	<0.01	0.09	.93
Time²	<0.01	-	<0.01	-0.71	.48
Spectral flux	-0.84	-0.35	0.17	-4.81	<.001
Positive affective state					
Intercept	2.09	-	0.13	16.64	<.001
Time	<-0.01	-	<0.01	-0.57	.57
Time²	<0.01	-	<0.01	0.71	.48
Spectral flux	1.42	0.45	0.23	6.28	<.001
Negative affective state					
Intercept	2.46	-	0.16	14.95	<.001
Time	<0.01	-	<0.01	1.56	.12
Time²	<0.01	-	<0.01	-1.99	.047
Spectral flux	-0.96	-0.24	0.25	-4.05	<.001
Valence					
Intercept	43.33	-	2.70	16.07	<.001
Time	<0.01	-	<0.01	1.61	.11
Time²	<0.01	-	<0.01	2.14	.03
Spectral flux	29.46	0.35	7.23	4.04	<.001
Energetic arousal					
Intercept	42.53	-	2.71	15.70	<.001
Time	<-0.01	-	<0.01	-3.14	.002
Time²	<0.01	-	<0.01	-3.96	<.001
Spectral flux	23.56	0.28	6.71	3.51	<.001
Calmness					
Intercept	40.58	-	3.39	11.98	<.001
Time	<0.01	-	<0.01	0.31	.75
Time²	<0.01	-	<0.01	2.99	.003
Spectral flux	23.81	0.26	6.86	3.47	<.001

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Table 16

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects for Spectral Flux of Voiced Regions Only, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.28	-	0.10	12.98	<.001
Time	<0.01	-	<0.01	0.07	.94
Time²	<0.01	-	<0.01	-0.58	.56
Spectral flux of voiced regions only	-0.55	-0.38	0.11	-4.81	<.001
Positive affective state					
Intercept	2.09	-	0.13	16.67	<.001
Time	<-0.01	-	<0.01	-0.54	.59
Time²	<0.01	-	<0.01	0.50	.62
Spectral flux of voiced regions only	0.89	0.46	0.15	5.94	<.001
Negative affective state					
Intercept	2.46	-	0.16	14.92	<.001
Time	<0.01	-	<0.01	1.53	.13
Time²	<0.01	-	<0.01	-1.82	0.07
Spectral flux of voiced regions only	-0.55	-0.11	0.16	-3.57	<.001
Valence					
Intercept	43.47	-	2.70	16.09	<.001
Time	<0.01	-	<0.01	1.63	.10
Time²	<0.01	-	<0.01	1.98	.048
Spectral flux of voiced regions only	16.91	0.32	4.80	3.52	<.001
Energetic arousal					
Intercept	42.57	-	2.71	15.70	<.001
Time	<-0.01	-	<0.01	-3.13	.002
Time²	<0.01	-	<0.01	-4.03	<.001
Spectral flux of voiced regions only	16.45	0.32	4.40	3.74	.027
Calmness					
Intercept	40.68	-	3.39	12.00	<.001
Time	<0.01	-	<0.01	0.33	.74
Time²	<0.01	-	<0.01	2.87	.004
Spectral flux of voiced regions only	14.23	0.26	4.51	3.16	.002

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Supplementary Materials Chapter 3

A3.1. ADS-K items with English translation in italics

1) Während der letzten Minuten haben mich Dinge beunruhigt, die mir sonst nichts ausmachen. | *During the last few minutes, things that normally don't bother me worried me.*

2) Während der letzten Minuten konnte ich meine trübsinnige Laune nicht loswerden, obwohl mich meine Freunde / Familie / Mitpatienten versuchten aufzumuntern. | *During the last few minutes, I couldn't get rid of my gloomy mood, although my friends / family / fellow patients tried to cheer me up.*

3) Während der letzten Minuten hatte ich Mühe mich zu konzentrieren: | *During the last few minutes I had trouble concentrating.*

4) Während der letzten Minuten war ich deprimiert / niedergeschlagen. | *During the last few minutes I was depressed / down.*

5) Während der letzten Minuten war alles anstrengend für mich. | *During the last minutes everything was exhausting for me.*

6) Während der letzten Minuten dachte ich, mein Leben ist ein einziger Fehlschlag. | *During the last minutes I thought my life was one big failure.*

7) Während der letzten Minuten hatte ich Angst. | *During the last minutes I was afraid.*

8) Während der letzten Minuten war ich fröhlich gestimmt. | *During the last minutes I was in a cheerful mood.*

9) Während der letzten Minuten habe ich weniger als sonst geredet.| *During the last minutes I talked less than usual.*

10) Während der letzten Minuten fühlte ich mich einsam. | *During the last minutes I felt lonely.*

11) Während der letzten Minuten habe ich das Leben genossen. | *During the last minutes I enjoyed life.*

12) Während der letzten Minuten war ich traurig. | *During the last minutes I felt sad.*

13) Während der letzten Minuten hatte ich das Gefühl, dass mich die Leute nicht leiden können. | *During the last minutes I felt that people didn't like me.*

14) Während der letzten Minuten konnte ich mich zu nichts aufraffen.
During the last minutes I couldn't get myself up to do anything.

Supplementary Materials Chapter 4

A4.1 ADS-K items with English translations in italics

1) Während der letzten Minuten haben mich Dinge beunruhigt, die mir sonst nichts ausmachen.
During the last few minutes, things that normally don't bother me worried me.

2) Während der letzten Minuten konnte ich meine trübsinnige Laune nicht loswerden, obwohl mich meine Freunde / Familie / Mitpatienten versuchten aufzumuntern.
During the last few minutes, I couldn't get rid of my gloomy mood, although my friends / family / fellow patients tried to cheer me up.

3) Während der letzten Minuten hatte ich Mühe mich zu konzentrieren.
During the last few minutes I had trouble concentrating.

4) Während der letzten Minuten war ich deprimiert / niedergeschlagen.
During the last few minutes I was depressed / down.

5) Während der letzten Minuten war alles anstrengend für mich.
During the last minutes everything was exhausting for me.

6) Während der letzten Minuten dachte ich, mein Leben ist ein einziger Fehlschlag.
During the last minutes I thought my life was one big failure.

7) Während der letzten Minuten hatte ich Angst.
During the last minutes I was afraid.

8) Während der letzten Minuten war ich fröhlich gestimmt.
During the last minutes I was in a cheerful mood.

9) Während der letzten Minuten habe ich weniger als sonst geredet.
During the last minutes I talked less than usual.

10) Während der letzten Minuten fühlte ich mich einsam.
During the last minutes I felt lonely.

11) Während der letzten Minuten habe ich das Leben genossen.
During the last minutes I enjoyed life.

12) Während der letzten Minuten war ich traurig.
During the last minutes I felt sad.

13) Während der letzten Minuten hatte ich das Gefühl, dass mich die Leute nicht leiden können.
During the last minutes I felt that people didn't like me.

14) Während der letzten Minuten konnte ich mich zu nichts aufraffen.
During the last minutes I couldn't get myself up to do anything.

A4.2 MDMQ items with English translations in italics

Im Moment fühle ich mich …	*At the moment I feel ...*
unzufrieden – zufrieden	*discontent- content*
unwohl – wohl	*unwell - well*
müde – wach	*tired - awake*

Im Moment fühle ich mich …	*At the moment I feel ...*
energielos – energiegeladen	*without energy - full of energy*
unruhig – ruhig	*agitated - calm*
angespannt – entspannt	*tense - relaxed*

A4.3 Positive and negative affect items with English translations in italics

Im Moment fühle ich mich

fröhlich / zufrieden / tatkräftig / enthusiastisch / entspannt / glücklich; einsam / traurig /

unsicher / ängstlich / niedergeschlagen / schuldig / deprimiert / misstrauisch / gereizt

At the moment I feel

cheerful / content / energetic / enthusiastic / relaxed / happy;

lonely, sad, insecure, anxious, depressed, low-spirited, guilty, distrustful, irritable

A4.4 Reasons for exclusion of selfie videos

- the full set of videos from four patients
  - one patient did not say anything during the videos (23 files)
  - in the videos of two patients no sound was recorded due to technical issues (30 files)
  - one patient provided only 2 videos (2)
- videos with technical damages (2)
- test runs (14) speech sample or the reported affective states were missing (19)
- files of consecutive assessments less than 15 minutes apart from each other (19); here only the first assessment was kept unless its audio quality was insufficient or only the second assessment included assessments of affective states; in such cases the second assessment was kept
-
- accidental recordings without content (30)
- files in which the microphone was masked (16)
- assessments in which either the

A4.5 Multilevel linear regression analysis

Table 17

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects of the Positive Emotion Words, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.27	-	0.10	12.93	<.001
Time	<0.01	-	<0.01	0.21	.83
Time²	<-0.01	-	<0.01	-0.20	.84
Positive emotion words	-0.02	-0.14	<0.01	-4.67	.84
Positive affective state					
Intercept	2.10	-	0.13	16.53	<.001
Time	<-0.01	-	<0.01	-0.79	.43
Time²	<-0.01	-	<0.01	-0.22	.83
Positive emotion words	0.03	0.16	<0.01	5.69	<.001
Negative affective state					
Intercept	2.46	-	0.17	14.73	<.001
Time	<0.01	-	<0.01	1.41	.16
Time²	<-0.01	-	<0.01	-1.73	.08
Positive emotion words	-0.02	-0.09	<0.01	-4.00	<.001
Valence					
Intercept	43.49	-	2.81	15.50	<.001
Time	0.17	-	0.13	1.32	.19
Time²	0.04	-	0.03	1.37	.17
Positive emotion words	1.12	0.22	0.18	6.09	<.001
Energetic arousal					
Intercept	42.58	-0.11	2.82	15.08	<.001
Time	-0.42	-0.16	0.12	-3.47	<.001
Time²	-0.12	0.20	0.02	-4.78	<.001
Positive emotion words	0.99	-0.11	0.17	5.91	<.001
Calmness					
Intercept	40.45	-	3.48	11.62	<.001
Time	<0.01	-	0.12	0.03	.99
Time²	0.07	0.09	0.02	2.65	.008
Positive emotion words	1.19	0.22	0.17	7.02	<.001

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Table 18

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects of Negative Emotion Words, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.27	-	0.10	12.95	<.001
Time	<0.01	-	<0.01	0.52	.60
Time²	<-0.01	-	<0.01	-0.15	.88
Negative emotion words	0.02	0.16	<0.01	6.13	<.001
Positive affective state					
Intercept	2.11	-	0.13		<.001
Time	<-0.01	-	<0.01	16.54	.26
Time²	<-0.01	-	<0.01	-1.14	.75
Negative emotion words	-0.03	-0.18	<0.01	-0.32	<.001
Negative affective state					
Intercept	2.46	-	0.17	14.73	<.001
Time	<0.01	-	<0.01	1.72	.085
Time²	<-0.01	-	<0.01	-1.70	.090
Negative emotion words	0.03	0.15	<0.01	5.69	<.001
Valence					
Intercept	43.57	-	2.81	15.52	<.001
Time	0.13	-	0.13	0.98	.33
Time²	0.03	-	0.03	1.26	.21
Negative emotion words	-1.24	-0.28	0.17	-7.31	<.001
Energetic arousal					
Intercept	42.64	-	2.83	15.09	<.001
Time	-0.46	-0.12	0.12	-3.85	<.001
Time²	-0.12	-0.16	0.02	-4.96	<.001
Negative emotion words	-1.11	-0.25	0.15	-7.15	<.001
Calmness					
Intercept	40.55	-	3.48	11.64	<.001
Time	-0.03	-	0.12	-0.25	.801
Time²	0.06	0.07	0.02	2.48	.014
Negative emotion words	-0.95	-0.20	0.16	-5.93	<.001

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Table 19

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects of First Person Singular Pronouns, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.27	-	0.10	12.85	<.001
Time	<0.01	-	<0.01	0.17	.87
Time²	<0.01	-	<0.01	0.05	.96
First person pronoun	<0.01	0.08	<0.01	1.84	.07
Positive affective state					
Intercept	2.11	-	0.13	16.56	<.001
Time	<-0.01	-	<0.01	-0.73	.47
Time²	<-0.01	-	<0.01	-0.48	.63
First person pronoun	<-0.01	-0.04	<0.01	-1.24	.21
Negative affective state					
Intercept	2.45	-	0.17	14.67	<.001
Time	<0.01	-	<0.01	1.36	.17
Time²	<0.01	-	<0.01	-1.43	.15
First person pronoun	0.01	0.07	<0.01	2.73	.007
Valence					
Intercept	43.66	-	2.80	15.59	<.001
Time	0.18	-	0.14	1.32	.19
Time²	0.03	-	0.03	1.13	.26
First person pronoun	-0.21	-0.05	0.17	-1.24	.22
Energetic arousal					
Intercept	42.71	-	2.83	15.10	<.001
Time	-0.42	-0.11	0.12	-3.36	<.001
Time²	-0.12	-0.16	0.02	-4.83	<.001
First person pronoun	-0.06	-0.01	0.15	-0.37	.71
Calmness					
Intercept	40.65	-	3.49	11.67	<.001
Time	<0.01	-	0.13	0.06	.95
Time²	0.06	0.07	0.03	2.32	.021
First person pronoun	-0.25	-0.05	0.16	-1.60	.11

Note. ADS-K = Allgemeine Depressionsskala Kurzform.

Table 20

Multilevel Linear Regression Analysis to Predict Momentary Depression Severity and Affective States: Fixed Effects of the Past Tense Words, Time, and Time²

Statistical predictor	Beta coefficient	Standardized beta coefficient	Standard error	T	*p*-value
ADS-K					
Intercept	1.27	-	0.10	12.87	<.001
Time	<0.01	-	<0.01	0.20	.84
Time²	<-0.01	-	<0.01	-0.03	.98
Past tense	<-0.01	-0.03	<0.01	-1.04	.30
Positive affective state					
Intercept	2.11	-	0.13	16.56	<.001
Time	<-0.01	-	<0.01	-0.77	.44
Time²	<-0.01	-	<0.01	-0.45	.65
Past tense	<0.01	0.05	<0.01	1.50	.13
Negative affective state					
Intercept	2.45	-	0.16	14.70	<.001
Time	<0.01	-	<0.01	1.37	.17
Time²	<-0.01	-	<0.01	-1.56	.12
Past tense	<-0.01	<-0.01	<0.01	-0.16	.88
Valence					
Intercept	43.62	-	2.80	15.58	<.001
Time	0.17	-	0.14	1.29	.20
Time²	0.03	-	0.03	1.18	.24
Past tense	0.21	0.04	0.20	1.04	.30
Energetic arousal					
Intercept	42.70	-	2.83	15.10	<.001
Time	-0.42	-0.11	0.12	-3.40	<.001
Time²	-0.12	-0.16	0.02	-4.83	<.001
Past tense	0.25	0.05	0.19	1.34	.18
Calmness					
Intercept	40.60	-	3.48	11.65	<.001
Time	<-0.01	-	0.13	<-0.01	.99
Time²	0.06	0.06	0.03	2.38	.017
Past tense	0.42	0.07	0.19	2.20	.028

Note. ADS-K = Allgemeine Depressionsskala Kurzform.