Variability of speech timing features across repeated recordings: a comparison of open-source extraction techniques

Judith Dineley[1], Ewan Carr[1], Lauren L. White[1], Catriona Lucas[1], Zahia Rahman[1], Tian Pan[1], Faith Matcham[2], Johnny Downs[1], Richard J. Dobson[1,3], Thomas F. Quatieri[4], Nicholas Cummins[1,5]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[2]School of Psychology, University of Sussex, Falmer, UK
[3]Institute of Health Informatics, University College London, London, UK
[4]MIT Lincoln Laboratory, Lexington, MA, USA
[5]thymia, London, UK

judith.dineley@kcl.ac.uk, nick.cummins@kcl.ac.uk

Abstract

Variations in speech timing features have been reliably linked to symptoms of various health conditions, demonstrating clinical potential. However, replication challenges hinder their translation; extracted speech features are susceptible to methodological variations in the recording and processing pipeline. Investigating this, we compared exemplar timing features extracted via three different techniques from recordings of healthy speech. Our results show that features extracted via an intensity-based method differ from those produced by forced alignment. Different extraction methods also led to differing estimates of within-speaker feature variability over time in an analysis of recordings repeated systematically over three sessions in one day (n=26) and in one week (n=28). Our findings highlight the importance of feature extraction in study design and interpretation, and the need for consistent, accurate extraction techniques for clinical research.

Index Terms: speech timing, feature extraction, reproducibility, longitudinal monitoring

1. Introduction

Speech offers an accessible, objective way to monitor an individual's health. It has the potential to measure outcomes in clinical trials and, ultimately, be applied in clinical practice [1]. To achieve this, analytical pipelines that reliably track clinical state over time are needed [2]. However, speech features are susceptible to multiple sources of variability along the extraction pipeline. Factors related to recording, such as room acoustics, and the speaker, such as time of day, voice use and menstruation, can produce non-pathological temporal variation in speech production [3-5].

Speaker, recording and extraction variability is problematic when using speech to assess health state: it can mask or even mimic pathological changes. For example, room reverberation [3] and recording earlier in the day ('morning voice') [4] can result in lower pitch that is also observed in speakers with major depression [6]. To develop reliable, interpretable prediction models based on these features, it is vital we characterize their variability and its effects [7].

The literature quantifying variability in speech feature extraction pipelines is sparse. Two studies have highlighted reliability concerns in acoustic features [4], [8]. However, neither study considered key timing features (e.g., average pause length) that contain important clinical information related to health conditions, including depression, amyotrophic lateral sclerosis (ALS) and Parkinson's disease [9-12].

Timing feature test-retest reliability was explored in 46 healthy adult American English speakers [13], with mixed results, finding moderate reliability for phrase duration but poor reliability for pause duration. Similarly, a study examining test-retest reliability in 40 healthy Chinese adults found approximately half of the features tested did not reach a moderate level of reliability [14]. In contrast to [13], the authors observed moderate reliability in mean pause duration [14]. Both papers omit to report certain methodological details such as the time of day of the recording, which, given the diurnal variations reported in [4], could have affected the findings. Further, the authors used different tools – intensity thresholding [13] and forced alignment [14] – to extract their timing features.

Fundamentally different approaches to extraction are likely to produce different speech timing features. Further, different extraction techniques may be more or less sensitive to variations in speech to due varying recording and speaker factors. Each approach requires different resources and expertise. The choice of extraction tool in speech-health research, therefore, is non-trivial. Manual transcription relies on the subjective identification of pause boundaries and is extremely labor-intensive [15]. Intensity thresholding can misidentify unvoiced speech segments as silence [15]. Meanwhile, extraction by identifying words and phonetic boundaries using transcription and forced alignment relies upon transcription performance; several transcription tools of varying accuracy exist [16].

Understanding the differences in extracted features that arise through differences in extraction pipeline design is key to developing reliable pipelines for clinical research and practice. Towards this aim, this study compares three established open-source timing feature extraction techniques. As a preliminary investigation, we analyzed healthy speech to understand the strengths and limitations of different extraction approaches and inform future studies of pathological speech. We used a previously described dataset of speech recorded repeatedly in a controlled environment that minimized variability attributable to hardware, set-up and acoustic conditions [17]. We describe differences in extracted timing features between techniques and investigate the influence of extraction methods on non-pathological within-speaker variability.

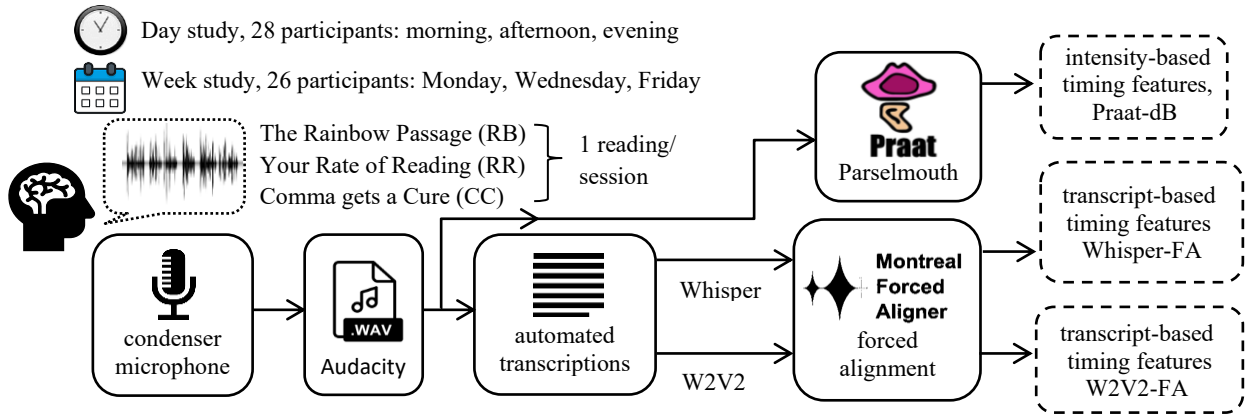

Figure 1: *Data recording and feature extraction pipelines. W2V2 – wave2vec automatic speech recognition (ASR) engine*

2. Methodology

We analyzed read speech recordings of 28 healthy participants speaking in the morning, afternoon and early evening (Day), and 26 participants speaking on a Monday, Wednesday, and Friday at the same time (Week) [17]. From these recordings, we extracted exemplar timing and pausing features using three methods (Figure 1). The first is a Praat script that uses intensity thresholding [18] The other two methods utilize word boundaries estimated from forced alignment, with transcripts generated using two different automatic speech recognition (ASR) engines: (i) Whisper (Open AI) [19], and (ii) wav2vec 2.0 (Meta) [20]. We refer to the three approaches herein as Praat-dB, Whisper-FA, and W2V2-FA, respectively.

We consider automated transcription methods as a realistic solution for a real-world digital health pipeline. In these two pipelines, we estimated word boundary estimates using the Montreal Forced Aligner (MFA) [21] rather than direct estimates from Whisper and W2V2. This allowed us to identify differences in these approaches attributable to transcription methods only, as the pipelines are otherwise identical.

We present three analyses. *Analysis 1* compares two transcript characteristics, word count and Levenshtein Distance (a character-level dissimilarity measure) [22] and how associated timing features, speaking rate and articulation rate, differ between the transcripts generated using Whisper-FA and W2V2-FA. *Analysis 2* compares variation in pause delineation between the three methods, highlighting the effects on five exemplar features commonly used in speech-health research, [9-14]: *phonation time ratio*, *pause time ratio*, *pause rate*, *mean pause duration* and the *coefficient of variation (CoV) of pause duration*. *Analysis 3* compares how the five features generated by the three methods are affected by within-speaker variation (over time). These features were selected to be independent of reading content.

2.1. Speech Dataset

Participant recruitment: 54 adults pre-screened to exclude any hearing, speaking, neurological or mental health disorders that might affect their speech were recruited (Table 1) [17].

Recording schedules: Participants in Day completed three recording sessions in a single day, scheduled in the morning (08:00-10:00), the afternoon (13:00-15:00), and the early evening (17:00-19:00). The minimum time between sessions was 3.5 hours. Week study participants were scheduled on each Monday, Wednesday, and Friday in a week at the same time each day to minimise the effect of within-day variability.

Recording room and setup: Speech was recorded in a room insulated for reverberation and with quiet surroundings. Recordings were made with an Audio Technica AT2020USB+ condenser microphone with no built-in pre-processing, operated using Audacity (v. 3.1.3).

Speech elicitation: Participants recited one of three texts per session to avoid practice effects that could add to timing feature variability: *The Rainbow Passage* (long version) (RB), *Your Rate of Oral Reading* (RR) and *Comma gets a Cure* (CC). Participants read RB in their first session and were randomly assigned either RR or CC in their second and third sessions. The three have similar lengths and structural and lexical complexity and are suitable for normative studies [23].

2.2. Feature Extraction

Our extraction pipelines used Audacity (.aup3) files converted to single-channel 16kHz, 16 bit, Waveform Audio File Format (WAV) files (Figure 1). The Praat-dB pipeline identifies pause boundaries using a preset intensity threshold [18]; it does not use transcriptions and, therefore, is not susceptible to recognition errors.

The Whisper-FA-derived features are generated from transcripts by the whisper-base.en model (Figure 1), followed by MFA alignment [21] using the provided English acoustic model (v2.0.0a). Whisper intentionally removes disfluencies and, therefore, does not provide verbatim transcripts [24], with potential implications for our timing feature estimates. The W2V2-FA extraction pipeline, in contrast, uses the *wav2vec2-base-960h* model and is designed to provide verbatim transcripts, albeit at the potential cost of a higher word error rate than Whisper [16]. For comparability, this pipeline also uses MFA to estimate word boundaries.

Software: All features were extracted in Python (v. 3.10.9). Praat-dB features were extracted using the Parselmouth Python library for running Praat (v. 0.4.3) [25]. Both ASRs were implemented using the Hugging Face Transformer API (v. 4.38) [26]. Code will be made available upon request.

Table 1: *Participant characteristics (n=54)* [17]. *IQR – interquartile range.*

		Day	Week
Sex	Female	15	17
	Male	13	9
English L1	Yes	24	17
	No	4	9
Age (years)	Median (IQR)	26 (23-34)	29 (24-34)

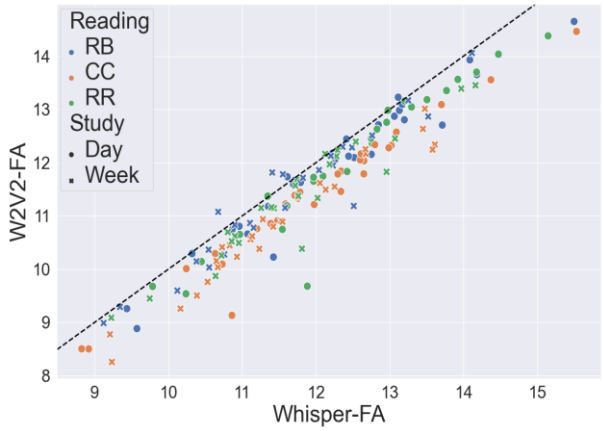

Figure 2: *A comparison of articulation rate (phones/s) extracted via Whisper-FA and W2V2-FA. Whisper-FA consistently estimates higher rates than W2V2-FA.*

2.3. Analysis

Analysis 1 and 2: We first calculated word counts and Levenshtein Distances to compare our two transcription-based pipelines. We then calculated pairwise differences between features generated from the different extraction methods. We used linear mixed effects (LME) models to estimate standardized mean differences between extraction methods for each feature (i.e., Praat-dB v Whisper-FA, Praat-dB vs W2V2-FA, Whisper-FA v W2V2-FA). Each LME model defined a single standardized speech feature as the outcome and included extraction method, sex, L1 English and study (i.e., *Day* or *Week*), plus two dummy variables indicating the recording session number. We report the β coefficients for extraction method and corresponding 95% confidence interval (CI). In this preliminary analysis, we focused on the direction of effects and their consistency when comparing methodologies.

Analysis 3: We calculated intraclass correlation coefficients (ICC) to explore the effect of non-pathological within-speaker variability on timing features derived from the three pipelines. The ICC metric is a measure of variance over repeated measurements, i.e., a feature extracted from recordings from different times of day and on days of the week. In this analysis, this metric indicates how much of the observed feature variability is attributable to within-participant (the same person, over a day or week) or between-participant differences. A value close to zero denotes observed feature variance is dominated by within-participant variation (over time); values close to one denote variation dominated by between-participant differences.

ICCs were calculated using LME models, with a separate model for each feature-study combination, e.g., one model for pause rate in *Day* data and another for pause rate in *Week* data. The LME models were unconditional and included the speech feature as the outcome, as in *Analysis 2*, and a participant-level random intercept only. If each pipeline were to extract timing features consistently from each recording, different pipelines should demonstrate similar within-individual variability in timing features. Our ICC definition differs to [13, 14] so we do not compare findings with those works.

Software: All tests were conducted in R (v. 4.3.2). *Analysis 1* and *2* were undertaken using the *lme4* package [27]. The bootstrap CIs were estimated using a parametric percentile bootstrap with 500 iterations, implemented using the *confint.merMod* method. *Analysis 3* was undertaken using the *glmmTMB* package [28]. Code is available upon request.

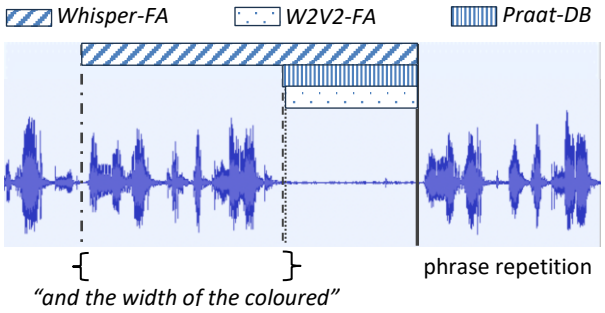

Figure 3: *Illustrative example comparing pause delineation, shown by the shaded bars, in which the speaker repeated a phrase. Praat-DB and W2V2-FA identified both instances. Whisper-FA transcribes the 2nd instance only, giving a markedly longer pause. Praat-DB estimated the start of the pause 40 ms earlier than W2V2-FA.*

3. Results and Discussion

The analysis dataset comprised 161 recordings of 54 participants (Table 1). One participant in *Week* missed one session out of three due to illness. Each participant recited each reading once. Mean duration of the recordings of each reading were 114s (RB), 103s (RR) and 130s (CC).

3.1. Analysis 1: Variation in transcription characteristics

Word count and Levenshtein Distance: Overall, the number of words transcribed were similar across ASR techniques (RB (session 1) mean word count (SD): *Day*: Whisper = 330 (3), W2V2 = 332 (7), *Week*: Whisper =330 (2), W2V2 = 333 (8)). We would not expect the speech of healthy participants reciting a set reading to be overly affected by dysfluencies, the transcription of which is a key difference between our ASR engines. The observed differences might arise from Whisper's text normalisation function improving transcipt readability. Levenshtein distances demonstrated differences between methods: RB (session 1) mean distance (SD), *Day*: 60 (1), *Week:* 68 (42). This is expected, as Whisper's text normalisation function changes characters to improve spelling and transcript readability.

LME Analysis: We observed no consistent difference in speaking rate between the transcription methods (mean standard difference (CI) = -0.02 (-0.07, 0.04)). In contrast, we observed a consistent difference in articulation rate (mean standard difference (CI) = 0.32 (0.26, 0.33)) (Figure 2), that may be due to character-level differences in transcripts as indicated by the Levenshtein Distance values.

A limitation of this analysis is the lack of manual transcriptions to compare with our automated pipelines. This highlights a barrier in extraction tool development: the large human resource needed to manually transcribe multiple recordings.

3.2. Analysis 2: Variation in pause delineation

A visual inspection of pause boundaries showed that Praat-dB tended to identify pauses as starting earlier and finishing later than Whisper-FA and W2V2-FA, e.g., RB (session 1), Day, mean pause duration (SD): Praat = 684 (140) ms, Whisper-FA = 663 (132) ms, W2V2-FA = 645 (118) ms. Similar differences were observed in the Week data. Whisper-FA and W2V2-FA typically differed where Whisper ASR removed repetitions and dysfluencies (Figure 3).

Table 2: *Standardised pairwise mean difference and confidence intervals (CI) using Linear Mixed Effect (LME) Models to account for repeated measures, adjusted for sex, L1 English, study (i.e., Day or Week), and recording session number (n = 54). In this preliminary analysis, we focus on the direction of effects and their consistency when comparing methodologies.*

	W2V2-FA – Whisper-FA		W2V2-FA – Praat-dB		Whisper-FA – Praat-dB	
Pause Duration (CoV)	0.23	[0.02, 0.45]	-0.28	[-0.48, -0.09]	-0.05	[-0.24, 0.13]
Pause Duration (Mean)	0.11	[-0.06, 0.27]	-0.32	[-0.44, -0.16]	-0.22	[-0.32, -0.12]
Pause Rate	0.12	[-0.05, 0.29]	-0.73	[-0.89, -0.57]	-0.64	[-0.76, -0.51]
Pause Time Ratio	0.16	[-0.01, 0.31]	-1.21	[-1.34, -1.09]	-1.22	[-1.22, -1.03]
Phonation Ratio	-0.14	[-0.33, 0.04]	0.83	[0.67, 0.98]	0.72	[0.59, 0.83]

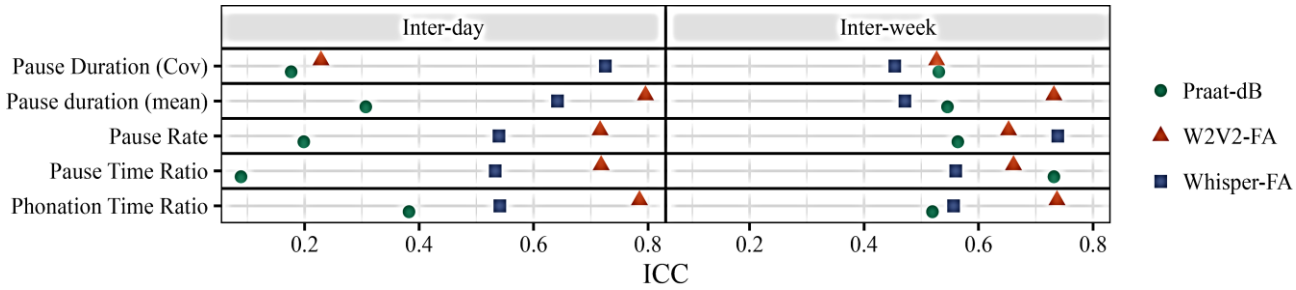

Figure 4: *ICC comparison of five speech timing features extracted using three pipelines from a set of three recordings of each speaker taken in one day (intra-day) or one week (intra-week). ICC values closer to zero indicate variability dominated by within-participant variability (i.e., between recordings of the same person). Values close to 1 indicate variability dominated by between-person variations.*

LME Analysis: We observed differences between almost all Praat-dB features and those from the two other pipelines (Table 2). We speculate this is due to small differences in individual pause boundaries resulting in larger cumulative differences over an entire recording. Other than pause duration CoV, our features generated using Whisper-FA and W2V2-FA do not appear to differ (Table 2). We speculate pause duration CoV differences could be due to variations introduced by Whisper's removal of repetitions and dysfluencies (Figure 3).

3.3. Analysis 3: Effect of non-pathological variability

Different extraction methods resulted in different ICC values (Figure 4). This indicates there are differences between extraction methods in the quantification of within-person feature variability over repeated feature measures. Feature extraction methods that extracted similar values would show similar feature variabilities over repeated recordings with similar ICC. Further, even if pipelines extracted different feature values from one another, if they did so systematically (e.g., mean pause duration is always 20ms longer using one method), we would also expect these to show similar variability across repeated recordings. We do not observe this. A deeper investigation of individual extraction techniques is needed to help understand these effects.

The range of ICC values is notably smaller in the *Week* study; Praat-dB has markedly lower ICCs for *Day* compared to *Week*, indicating greater within-person variability. We speculate that as an intensity-based extraction approach, Praat-dB may be more sensitive to non-pathological variation in participants' voice intensity, resulting in greater within-individual variability with this approach.

W2F2-FA tended to exhibit lower within-person variability (higher ICC) than Whisper-FA. As the only difference between the two is the transcription methods, this must be attributable to differences in transcription. We propose that Whisper's lower ICC values are attributable to the extraction pipeline's insensitivity to variations in speech between participants.

Additional research, including on other datasets, will be needed to validate these findings. However, this exploration highlights the variability in speech features introduced by different feature extraction approaches. A similar analysis within clinical populations is needed to assess the impact of this finding in digital health systems.

4. Conclusion

Our exploratory analysis of healthy speech recorded with minimal variation in set-up and acoustic conditions demonstrated consistent differences between timing feature extraction pipelines. To predict health state using speech, differences in extraction pipelines must be understood and accounted for, particularly in clinical cohorts where differences attributable to extraction methods may be larger. For example, transcription tools that deliberately remove dysfluencies should be used with caution, as dysfluencies are more common in pathological speech and may also contain valuable health information. Differences in forced alignment performance additionally have implications for acoustic feature extraction, which we will investigate in future work.

Using two sets of repeated recordings of participants that adhered to a strict schedule of sessions in the morning, afternoon and evening, and three days in one week at the same time, variations in ICC between pipelines suggest that individual feature extraction pipelines do not extract features consistently over repeated recordings. Timing feature extraction that performs consistently over repeated recordings without losing salient information is critical for robust, informative speech tools [1, 2]. Though further investigation is needed to understand the source of inconsistencies, our findings point to opportunities to develop feature extraction pipelines with more consistent performance.

5. Acknowledgements

We thank our participants for their support and the KCL Department of Psychology for use of their test rooms. This work was supported by the EPSRC UK Acoustics Network Plus Pilot Fund and an IPEM Innovation grant. This paper also represents independent research part funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London and supported by the National Institute for Health and Care Research University College London Hospitals Biomedical Research Centre. Views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Dept. of Health and Social Care. For T.F. Quatieri: Material is approved for public release, distribution is unlimited, and is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

6. References

[1] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. Green, "Speech as a Biomarker: Opportunities, Interpretability, and Challenges," *Perspect. ASHA Spec. Interest Groups*, vol. 7, no. 1, pp. 276–283, Feb. 2022, doi: 10.1044/2021_PERSP-21-00174.

[2] A. P. Vogel and P. Maruff, "Monitoring change requires a rethink of assessment practices in voice and speech," *Logoped. Phoniatr. Vocol.*, vol. 39, no. 2, pp. 56–61, Jul. 2014, doi: 10.3109/14015439.2013.775332.

[3] J. Dineley et al., "Towards robust paralinguistic assessment for real-world mobile health (mHealth) monitoring: an initial study of reverberation effects on speech," in *Proc. INTERSPEECH 2023*, Dublin, Ireland: ISCA, 2023, pp. 2373–2377. doi: 10.21437/Interspeech.2023-947.

[4] J. L. Pierce, K. Tanner, R. M. Merrill, L. Shnowske, and N. Roy, "Acoustic Variability in the Healthy Female Voice Within and Across Days: How Much and Why?" *J. Speech Lang. Hear. Res.*, vol. 64, no. 8, pp. 3015–3031, Aug. 2021, doi: 10.1044/2021_JSLHR-21-00018.

[5] C. Ge, Y. Xiong, and P. Mok, "How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements," *Proc. INTERSPEECH 2021*. ISCA, Brno, Czech Republic, pp. 1683–1687, 2021. doi: 10.21437/Interspeech.2021-1122.

[6] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investig. Otolaryngol.*, vol. 5, no. 1, pp. 96–116, 2020, doi: 10.1002/lio2.354.

[7] M. E. McNamara, M. Zisser, C. G. Beevers, and J. Shumake, "Not just 'big' data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions," *Behav. Res. Ther.*, vol. 153, p. 104086, 2022, doi: 10.1016/j.brat.2022.104086.

[8] G. M. Stegmann et al., "Repeatability of Commonly Used Speech and Language Features for Clinical Applications," *Digit. Biomark.*, vol. 4, no. 3, pp. 109–122, Dec. 2020, doi: 10.1159/000511671.

[9] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," *Biol. Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012, doi: 10.1016/j.biopsych.2012.03.015.

[10] N. Cummins et al., "Multilingual markers of depression in remotely collected speech samples," *J. Affect. Disord.*, vol. 341, pp. 128–136, 2023, doi: 10.1016/j.jad.2023.08.097.

[11] J. R. Green et al., "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotroph. Lateral Scler. Frontotemporal Degener.*, vol. 14, no. 7–8, pp. 494–500, Dec. 2013, doi: 10.3109/21678421.2013.817585.

[12] S. Skodda, "Aspects of speech rate and regularity in Parkinson's disease," *J. Neurol. Sci.*, vol. 310, no. 1, pp. 231–236, 2011, doi: 10.1016/j.jns.2011.07.020.

[13] C. Barnett et al., "Reliability and validity of speech & pause measures during passage reading in ALS," *Amyotroph. Lateral Scler. Frontotemporal Degener.*, vol. 21, no. 1–2, pp. 42–50, Jan. 2020, doi: 10.1080/21678421.2019.1697888.

[14] F. Feng et al., "Test-retest reliability of acoustic and linguistic measures of speech tasks," *Comput. Speech. Lang.*, vol. 83, p. 101547, 2024, doi: 10.1016/j.csl.2023.101547.

[15] J. R. Green, D. R. Beukelman, and L. J. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech," *J. Med. Speech. Lang. Pathol.*, vol. 12, no. 4, p. 149, 2004.

[16] A. Romana, K. Koishida, and E. Mower Provost, "Automatic Disfluency Detection from Untranscribed Speech," *arXiv*. 2023. doi: 10.48550/arXiv.2311.00867.

[17] N. Cummins et al., "A pilot protocol and cohort for the investigation of non-pathological variability in speech." *arXiv*, 2024. Accessed: Jun. 12, 2024. [Online]. Available: https://arxiv.org/abs/2406.07497

[18] N. H. de Jong, J. Pacilly, and W. Heeren, "PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically," *Assess. Educ*., vol. 28, no. 4, pp. 456–476, Jul. 2021, doi: 10.1080/0969594X.2021.1951162.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in 40th International Conference on Machine Learning, Vienna, Austria: PMLR, 2023, pp. 28492–28518.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, Vancouver, Canada: Curran Associates, 2020, pp. 12449–12460.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Proc. INTERSPEECH 2017*, Stockholm, Sweden: ISCA, 2017, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.

[22] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Sov. Phys.-Dokl*, vol 10, no. 8, 1966, pp. 707–710.

[23] T. W. Powell, "A comparison of English reading passages for elicitation of speech samples from clinical populations," *Clin. Linguist. Phon.*, vol. 20, no. 2–3, pp. 91–97, 2006, doi: 10.1080/02699200400026488.

[24] C. Lea et al., "From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition," in *Conference on Human Factors in Computing Systems*, Hamburg, Germany: ACM, 2023. doi: 10.1145/3544548.3581224.

[25] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phon.*, vol. 71, pp. 1–15, 2018, doi: 10.1016/j.wocn.2018.07.001.

[26] T. Wolf et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv*. 2019. doi: 10.48550/arXiv.1910.03771.

[27] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[28] M. E. Brooks et al., "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *R. J.*, vol. 9, no. 2, pp. 378–400, 2017, doi: 10.32614/RJ-2017-066.