Voice Stress Markers Are Orthogonal to Speech Disfluency Labels: A Large-Scale Analysis on SEP-28K

Nazar Kozak
Kozak Technologies Inc., Los Angeles, CA, USA
nzrkzk@gmail.com
ORCID: 0009-0001-8858-6098

April 2026

Abstract

The relationship between voice stress markers and speech disfluency events has not been systematically quantified at scale, despite both being targets of clinical assessment in stuttering populations. We examine correlations between four acoustic stress features—jitter, shimmer, fundamental frequency (F0) standard deviation, and a composite stress score—and five disfluency types (prolongation, block, sound repetition, word repetition, interjection) across 14,645 three-second clips from the SEP-28K dataset with valid pitch estimates. Using both Pearson and point-biserial correlations with Bonferroni correction for 20 comparisons, we find that all absolute correlations fall below 0.05, with all effect sizes negligible by Cohen's convention ($|r| < 0.10$). The strongest observed association (composite stress $\times$ prolongation, $r = -0.050$) explains only 0.25% of variance. Distribution comparisons between fluent and disfluent clips yield Cohen's $d < 0.10$ for all stress features. These findings suggest that, at least in terms of linear associations in this dataset, acoustic voice stress markers and disfluency labels carry largely non-overlapping information. While nonlinear or conditional dependencies cannot be ruled out from marginal correlations alone, the negligible effect sizes suggest that multimodal speech assessment systems may benefit from treating disfluency detection and stress monitoring as separate modules rather than modeling them jointly. We release analysis code and detailed statistical outputs to support reproducibility.

Keywords: voice stress analysis, disfluency detection, stuttering, SEP-28K, orthogonality, clinical speech assessment, jitter, shimmer, F0 variability

1 Introduction

Speech disfluency—encompassing prolongations, blocks, sound and word repetitions, and interjections—affects approximately 5% of children during developmental years, with about 1% persisting into adulthood (Yairi & Ambrose, 2013). Speech-language pathologists (SLPs) routinely assess both the overt speech disruptions (using instruments such as SSI-4 (Riley, 2009)) and the speaker's psychological state, including communication anxiety and voice tension (Yaruss & Quesal, 2006).

A growing body of work aims to automate stuttering detection using machine learning (Lea et al., 2021; Sheikh et al., 2022; Bayerl et al., 2022, 2023). Separately, voice stress analysis—extracting jitter, shimmer, and F0 variability as indicators of psychological arousal—has a long history in affective computing (Scherer, 2003; Cummins et al., 2015). However, the empirical relationship between these two signal dimensions in stuttering populations has not been systematically quantified at scale.

We address this gap by computing correlations between four voice stress features and five disfluency types across 14,645 clips from the SEP-28K dataset (Lea et al., 2021). Our key finding is that all linear correlations between voice stress markers and disfluency labels are negligible in magnitude ($|r| < 0.05$, Cohen's $d < 0.04$).

1.1 Contributions

1. We present the first large-scale quantitative analysis of the relationship between acoustic voice stress markers and per-type disfluency labels in a stuttering population ($N = 14{,}645$).
2. We report that all 20 stress–disfluency correlations are negligible ($|r| < 0.05$, Cohen's $d < 0.04$), with Bonferroni correction applied, indicating a lack of practically meaningful linear associations.
3. We discuss practical implications of this finding for multimodal speech analysis systems.
4. We release analysis code and statistical outputs for reproducibility.

2 Related Work

2.1 Acoustic Correlates of Stuttering

Research on the acoustic properties of stuttered speech has primarily focused on characterizing disfluency events themselves—their duration, spectral properties, and temporal patterns (Howell, 2004). Less attention has been paid to whether global voice quality measures differ between fluent and disfluent segments within the same speaker.

Goberman et al. (2010) examined correlations between stuttering severity and selected acoustic measures in adults who stutter, reporting that F0 variability was not consistently related to stuttering frequency. However, this study used severity ratings rather than per-event disfluency labels, and examined correlations at the speaker level rather than the clip level.

2.2 Voice Stress Analysis

Acoustic indicators of psychological stress include increased jitter (cycle-to-cycle pitch perturbation), increased shimmer (amplitude perturbation), elevated F0 mean and variability, and changes in speech rate (Scherer, 2003; Pisanski et al., 2016). These features have been used in affective computing for stress detection (Cummins et al., 2015), deception detection, and clinical anxiety assessment.

An implicit assumption in some multimodal speech systems is that stress markers may covary with disfluency—i.e., that speakers are more stressed when stuttering, and that this stress is acoustically detectable. Our analysis provides the first large-scale empirical test of this assumption.

2.3 SEP-28K Dataset

The SEP-28K dataset (Lea et al., 2021) contains 28,177 three-second clips extracted from podcasts featuring people who stutter, annotated by three raters for five disfluency types. Bayerl et al. (2022) demonstrated that four hosts account for 59% of clips and that evaluation protocol (random vs. speaker-independent splits) substantially affects reported performance. We use this dataset because it provides per-clip, per-type disfluency labels—essential for our correlation analysis.

3 Methods

3.1 Data Preparation

From SEP-28K, we retained 20,131 clips after filtering for audio quality, speech presence, and episode availability (258 of 385 episodes). Binary disfluency labels were established by majority vote ($\geq 2/3$ annotators). Table 1 summarizes the dataset.

Table 1: Dataset statistics after quality filtering ($N = 20{,}131$).

Disfluency Type	Positive Clips	Prevalence
Prolongation	2,001	9.9%
Block	2,469	12.3%
Sound Repetition	1,796	8.9%
Word Repetition	2,308	11.5%
Interjection	4,681	23.3%
Any disfluency	10,724	53.3%
Fluent	9,407	46.7%

3.2 Voice Stress Feature Extraction

Each 16 kHz mono clip was processed through a voice stress analysis pipeline extracting four features:

1. **Jitter (%)**: Cycle-to-cycle variation in pitch period, computed as the mean absolute difference between consecutive F0 periods divided by the mean period.
2. **Shimmer (%)**: Cycle-to-cycle variation in amplitude, computed as the mean absolute difference between consecutive frame amplitudes divided by the mean amplitude.
3. **F0 Standard Deviation (Hz)**: Standard deviation of fundamental frequency across voiced frames within the clip.
4. **Composite Stress Score**: Weighted combination: jitter (30%) + F0 variance (30%) + shimmer (20%) + speech rate deviation (20%), with per-speaker adaptive normalization.

3.2.1 Pitch Detection

Fundamental frequency (F0) was extracted using the YIN algorithm (de Cheveigné & Kawahara, 2002), which computes a cumulative mean normalized difference function (CMNDF). Implementation parameters: frame size 1024 samples (64 ms at 16 kHz), hop size 512 samples (32 ms), CMNDF threshold 0.15, with a global minimum fallback at threshold 0.4. This achieves 98.1% F0 detection rate across SEP-28K clips with speech content, compared to 12.1% with naive autocorrelation.

3.2.2 Adaptive Baseline

To account for individual vocal differences, we employ per-speaker adaptive baseline calibration using Welford's online algorithm (Welford, 1962):

- **Calibration phase**: Running mean and variance computed over the first 10 voiced frames ($\sim$10–20 seconds of speech).

- **Tracking phase**: Exponential moving average with $\alpha = 0.02$ for drift adaptation.
- **Scoring**: Each feature is z-scored against the adaptive baseline, transformed via sigmoid (steepness 1.5), and combined as a weighted composite.

This adaptive approach corrects for systematic voice quality differences between speakers. Without it, speakers with naturally high jitter or variable F0 would be incorrectly classified as stressed.

3.3 Analysis Subset

Of the 20,131 clips, 14,645 (72.7%) yielded valid F0 estimates (at least one voiced frame with detectable pitch). The remaining clips contained silence, noise, or unvoiced speech. All subsequent analyses use this subset of $N = 14{,}645$.

3.4 Statistical Analysis

We computed two types of correlations between each stress feature and each disfluency label:

1. **Pearson correlation** (r): Measures linear association between the continuous stress feature and the binary (0/1) disfluency label.
2. **Point-biserial correlation** (r_{pb}): The appropriate correlation for a continuous variable with a dichotomous variable; mathematically equivalent to Pearson r when one variable is binary.

With 4 stress features $\times$ 5 disfluency types = 20 comparisons, we applied Bonferroni correction ($\alpha_{\text{adjusted}} = 0.05/20 = 0.0025$) to control the family-wise error rate.

Effect sizes were interpreted using Cohen's conventions: negligible ($|r| < 0.10$), small ($0.10 \leq |r| < 0.30$), medium ($0.30 \leq |r| < 0.50$), and large ($|r| \geq 0.50$).

For distribution comparisons, we computed Cohen's d between fluent and disfluent clips for each stress feature.

4 Results

4.1 Correlation Analysis

Table 2 presents Pearson correlations between all stress features and disfluency labels.

Table 2: Pearson correlation (r) between voice stress features and disfluency labels ($N = 14{,}645$). Superscript * indicates significance after Bonferroni correction ($\alpha = 0.05/20$).

Feature	Prolong.	Block	SoundRep	WordRep	Interj.
Jitter (%)	$-.028^*$	$-.006$	.008	.013	.001
Shimmer (%)	$-.030^*$	.002	.019	.013	$-.002$
F0 Std Dev	$-.028^*$	$-.007$	$-.009$	.021	.001
Stress Score	$-.050^*$	$-.013$	$-.007$	.025	.004

Key observations:

- **Maximum absolute correlation**: $|r| = 0.050$ (composite stress $\times$ prolongation).
- **Mean absolute correlation**: $|r| = 0.015$ across all 20 pairs.

- **All effect sizes**: Negligible by Cohen's convention ($|r| < 0.10$).
- **Variance explained**: The strongest correlation explains only $r^2 = 0.0025$ (0.25%) of variance.
- **Statistical significance**: Four prolongation correlations reach significance after Bonferroni correction due to the large sample size, but their practical magnitude is negligible.

Point-biserial correlations (not shown) produced identical values, confirming the mathematical equivalence when one variable is binary.

4.2 Distribution Comparisons

We compared stress feature distributions between fluent ($N = 6{,}852$) and disfluent ($N = 7{,}793$) clips (subsets of the 14,645 with valid F0).

Table 3: Cohen's d effect sizes: fluent vs. disfluent clips.

Feature	Fluent Mean	Disfluent Mean	Diff	Cohen's d
Jitter (%)	12.365	12.357	−0.009	−0.001
Shimmer (%)	26.513	26.348	−0.165	−0.010
F0 Std Dev	55.375	55.151	−0.224	−0.006
Stress Score	0.734	0.728	−0.006	−0.034

All Cohen's d values are below 0.04 in absolute terms, well below the negligible threshold of 0.10. The distributions of voice stress features are virtually identical between fluent and disfluent clips.

4.3 Per-Type Analysis

We also examined whether specific disfluency types show stronger stress associations than others. Prolongation shows the largest (though still negligible) correlations across all four stress features, with a consistent negative direction. This could reflect a subtle acoustic artifact: prolonged sounds may reduce local jitter and shimmer measurements due to their sustained, quasi-periodic nature. Blocks, repetitions, and interjections show no systematic pattern.

5 Discussion

5.1 Lack of Linear Association Between Stress and Disfluency

Our central finding is that acoustic voice stress markers and disfluency labels show no practically meaningful linear correlations in SEP-28K. This is not merely a statistical null result—it is informative for system design.

The practical implication is that **voice stress features are poor linear predictors of disfluency labels, and vice versa**, at least in this dataset. This is consistent with treating stress assessment and disfluency detection as separable dimensions in multimodal systems, though we cannot rule out non-linear or context-dependent relationships from marginal correlations alone.

5.2 Practical Implications

The negligible linear associations observed here are consistent with treating disfluency detection and stress assessment as separable tasks in multimodal speech analysis. If confirmed in other datasets and populations, this would support modular system designs where disfluency and

stress modules operate independently, and stress measurements taken during disfluent speech are not confounded by the disfluency itself. However, we emphasize that this is a single-dataset observation on adult podcast speakers, and generalization requires further study.

5.3 Significance Despite Null Correlation

Large-sample null findings are scientifically valuable but underreported due to publication bias. Our result, with $N = 14{,}645$ and 20 controlled comparisons, provides strong evidence against a meaningful linear relationship. The 95% confidence interval for the largest correlation ($r = -0.050$) is approximately $[-0.066, -0.034]$, meaning that even under the most generous interpretation, the true population correlation is bounded well below the "small effect" threshold of $|r| = 0.10$.

5.4 Limitations

1. **Adult speakers only**: SEP-28K contains adult podcast hosts. Children's voice characteristics differ substantially (F0 range 250–400 Hz vs. adult male 85–180 Hz), and the stress–disfluency relationship may differ in pediatric populations.

2. **Linear analysis**: We examined linear (Pearson) correlations. Non-linear or conditional relationships (e.g., stress predicts disfluency only above a threshold) cannot be ruled out.

3. **Podcast context**: Podcast speakers may have adapted their speaking patterns for a public audience, potentially reducing observable stress.

4. **Per-clip analysis**: We correlate clip-level features with clip-level labels. Speaker-level aggregation might reveal different patterns, though our large N partially mitigates this.

5. **Acoustic stress only**: We measure acoustic correlates of stress, not physiological stress directly. Future work should incorporate heart rate, HRV, and electrodermal activity.

5.5 Future Directions

1. **Pediatric validation**: Replicate this analysis on FluencyBank (MacWhinney, 2024) or similar pediatric datasets.

2. **Non-linear analysis**: Apply mutual information or neural-network-based dependency measures.

3. **Physiological integration**: Combine acoustic stress with heart rate variability, electrodermal activity, and motion sensors.

4. **Longitudinal analysis**: Examine whether the stress–disfluency relationship changes within a speaking session (e.g., stress decreases as the speaker warms up, while disfluency frequency may change independently).

6 Conclusion

We presented a large-scale analysis ($N = 14{,}645$) examining linear associations between acoustic voice stress markers (jitter, shimmer, F0 variability, composite stress score) and speech disfluency labels in the SEP-28K dataset. All 20 correlations are negligible ($|r| < 0.05$, Cohen's $d < 0.04$), with the strongest explaining only 0.25% of variance. These results are consistent with the architectural choice of separating disfluency detection from stress assessment in clinical speech systems, as the two signal dimensions show negligible linear overlap in this dataset. Non-linear or population-specific dependencies remain to be investigated.

Data Availability

The SEP-28K dataset is publicly available (Lea et al., 2021). Analysis code (Python scripts for stress feature extraction and correlation analysis) will be released at `https://github.com/AnonAuthor/stress-disfluency-analysis` upon publication.

Acknowledgments

The preparation of this manuscript was assisted by Claude (Anthropic), an AI language model, used for drafting and editing. All technical content, analysis, and experimental design were conducted by the author.

References

Bayerl, S. P. et al. (2022). The influence of dataset partitioning on dysfluency detection systems. In *Proc. ACL Workshop on Computational Approaches to Linguistic Code-Switching*.

Bayerl, S. P. et al. (2023). Cross-corpus stuttering detection as a multi-label problem. In *Proc. Interspeech*. ISCA.

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.

de Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4), 1917–1930.

Goberman, A. M., Hughes, S., & Haydock, T. (2010). Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech Communication*, 52(10), 867–876.

Howell, P. (2004). Assessment of some contemporary theories of stuttering that apply to spontaneous speech. *Contemporary Issues in Communication Science and Disorders*, 31, 122–139.

Lea, C., Mitra, V., Joshi, A., Kajarekar, S., & Bigham, J. P. (2021). SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter. In *Proc. ICASSP*, 6798–6802. IEEE.

MacWhinney, B. (2024). FluencyBank: A TalkBank corpus for fluency disorders research. `https://fluency.talkbank.org/`.

Pisanski, K. et al. (2016). Voice parameters predict sex-specific body morphology in men and women. *Animal Behaviour*, 112, 13–22.

Riley, G. D. (2009). *Stuttering Severity Instrument–Fourth Edition (SSI-4)*. Pro-Ed.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2), 227–256.

Sheikh, S. A. et al. (2022). Machine learning for stuttering identification: Review, challenges, and future directions. *Neurocomputing*, 514, 385–402.

Welford, B. P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3), 419–420.

Yairi, E. & Ambrose, N. (2013). Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, 38(2), 66–87.

Yaruss, J. S. & Quesal, R. W. (2006). Overall assessment of the speaker's experience of stuttering (OASES). *Journal of Fluency Disorders*, 31(2), 90–115.