Adaptive Baseline Calibration for Voice Stress Assessment in Speech Disfluency Monitoring

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

April 2026

Abstract

Voice stress assessment systems commonly employ fixed thresholds for classifying acoustic features (jitter, shimmer, F0 variability) into stress levels. We show that fixed thresholds produce highly skewed stress score distributions when applied to diverse speakers, with 61.4% of clips scored as high-stress (≥ 0.8) in the SEP-28K dataset—likely an artifact of inter-speaker vocal variability rather than genuine stress variation, given the informal podcast recording context. We propose an adaptive baseline algorithm using Welford's online algorithm for per-speaker calibration, followed by exponential moving average tracking. Applied to 14,645 clips with valid pitch estimates, the adaptive approach produces a more symmetric distribution ($\mu = 0.530$, $\sigma = 0.162$) with substantially fewer extreme scores. We note that in the absence of ground-truth stress labels, we evaluate calibration quality by distribution shape rather than classification accuracy—a limitation shared by most voice stress analysis systems. We additionally report that YIN-based pitch detection achieves 98.1% F0 extraction rate on SEP-28K, compared to 12.1% with naive autocorrelation—a prerequisite for reliable voice stress features. We discuss implications for pediatric speech applications, where children's vocal characteristics (F0 range 250–400 Hz) differ substantially from adults and make fixed thresholds particularly problematic. The adaptive baseline algorithm is implemented in DisfluoSDK, an on-device framework for speech disfluency monitoring.

Keywords: voice stress analysis, adaptive baseline, Welford's algorithm, pitch detection, YIN, pediatric speech, speaker normalization, F0 variability

1 Introduction

Voice stress analysis—the extraction and classification of acoustic features indicative of psychological arousal—is a component of multimodal speech assessment systems for clinical populations including people who stutter (Scherer, 2003; Cummins et al., 2015). Features such as jitter (pitch perturbation), shimmer (amplitude perturbation), and fundamental frequency (F0) variability have been associated with emotional stress and anxiety (Pisanski et al., 2016).

A fundamental challenge in voice stress assessment is that the same acoustic feature values can indicate different stress levels for different speakers. A speaker with a naturally breathy voice may have high baseline shimmer; a speaker with a naturally variable F0 may have high F0 standard deviation even when calm. Fixed thresholds—the standard approach—cannot account for this variability.

This problem is particularly acute in pediatric populations, where:

- Children's F0 ranges (250–400 Hz for ages 5–10) are substantially higher than adult males (85–180 Hz) (Lee et al., 1999).

- Children's jitter and shimmer values tend to be higher due to incomplete laryngeal control (Nicollas et al., 2008).
- Inter-speaker variability is greater in children than in adults.

We present an adaptive baseline algorithm that calibrates voice stress scoring to each individual speaker and compare it with fixed thresholds on the SEP-28K dataset (Lea et al., 2021). We discuss its implications for pediatric speech monitoring applications.

1.1 Contributions

1. We quantify the apparent overestimation produced by fixed thresholds for voice stress assessment (61.4% of clips scored as high-stress on SEP-28K, where the informal podcast context makes widespread genuine stress implausible).
2. We propose a Welford-based adaptive baseline algorithm with EMA tracking that produces a near-symmetric stress score distribution ($\mu = 0.530$, $\sigma = 0.162$), consistent with more plausible calibration under heterogeneous speakers.
3. We demonstrate that YIN-based pitch detection achieves 98.1% F0 extraction rate vs. 12.1% for naive autocorrelation.
4. We define age-adjusted normalization ranges for pediatric voice stress assessment across four age groups.

2 Related Work

2.1 Voice Stress Analysis

Scherer (2003) provided a comprehensive review of vocal communication of emotion, establishing that F0 level, F0 variability, speech rate, and voice quality are modulated by emotional state. Cummins et al. (2015) reviewed speech analysis for depression and suicide risk, demonstrating clinical utility of acoustic features but also highlighting the need for speaker normalization.

2.2 Speaker Normalization

Speaker normalization techniques in ASR (vocal tract length normalization, cepstral mean normalization) are well established (Lee & Rose, 1996). However, these techniques are designed for improving recognition accuracy, not for stress-level assessment. Stress assessment requires normalizing the *magnitude* of features to a speaker-specific baseline while preserving *deviations* that indicate state changes.

2.3 Pitch Detection

F0 estimation is the foundation of voice stress features. de Cheveigné & Kawahara (2002) proposed the YIN algorithm, which uses a cumulative mean normalized difference function (CMNDF) to achieve robust pitch detection. Modern implementations typically use autocorrelation-based methods, but naive implementations suffer from subharmonic errors and low detection rates.

3 Methods

3.1 Dataset

We use 20,131 clips from the SEP-28K dataset (Lea et al., 2021) after quality filtering. Of these, 14,645 (72.7%) yield valid F0 estimates. Each clip is 3 seconds of 16 kHz mono audio from podcasts by people who stutter.

3.2 Pitch Detection: YIN Implementation

Our YIN implementation uses the following parameters:

- Frame size: 1024 samples (64 ms at 16 kHz)
- Hop size: 512 samples (32 ms)
- CMNDF threshold: $\tau_{th} = 0.15$
- Global minimum fallback: $\tau_{gm} = 0.4$
- F0 range: 75–500 Hz (adult), 150–500 Hz (pediatric)
- RMS voiced speech threshold: 0.005

The implementation leverages Apple's Accelerate framework (vDSP) for efficient computation of the autocorrelation and difference functions.

3.3 Voice Stress Features

Four features are extracted per clip:

1. **Jitter (%)**: $\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \,/\, \bar{T}$, where T_i are F0 periods.
2. **Shimmer (%)**: $\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}| \,/\, \bar{A}$, where A_i are frame amplitudes.
3. **F0 Std Dev (Hz)**: Standard deviation of F0 across voiced frames.
4. **Composite Stress Score**: Weighted combination with adaptive normalization.

3.4 Fixed Threshold Approach (Baseline)

The standard fixed-threshold approach applies population-level cutoffs:

- Jitter: calm $< 1\%$, stressed $> 3\%$
- Shimmer: calm $< 3\%$, stressed $> 8\%$
- F0 Std Dev: calm < 15 Hz, stressed > 40 Hz

Each feature is linearly mapped to $[0, 1]$ between the calm and stressed thresholds, then combined as:

$$S_{\text{fixed}} = 0.30 \cdot s_{\text{jitter}} + 0.30 \cdot s_{\text{F0var}} + 0.20 \cdot s_{\text{shimmer}} + 0.20 \cdot s_{\text{rate}}$$

Algorithm 1 Adaptive Baseline Stress Scoring

Require: Stream of voiced frames $\{f_1, f_2, \ldots\}$

Initialize Welford accumulators for each feature
$n_{\text{cal}} \leftarrow 0$; confidence $\leftarrow 0$
for each voiced frame f_t **do**
  Extract features: $x_t = [\text{jitter}_t, \text{shimmer}_t, \text{F0std}_t, \text{rate}_t]$
  if $n_{\text{cal}} < 10$ **then**
    Calibration: Update Welford mean μ and variance σ^2 for each feature
    $n_{\text{cal}} \leftarrow n_{\text{cal}} + 1$
    confidence $\leftarrow n_{\text{cal}}/10$
  else
    Tracking: $\mu \leftarrow \alpha \cdot x_t + (1 - \alpha) \cdot \mu \quad (\alpha = 0.02)$
  end if
  Z-score: $z_t^{(k)} = (x_t^{(k)} - \mu^{(k)})/\sigma^{(k)}$ for each feature k
  Sigmoid: $s_t^{(k)} = 1/(1 + e^{-1.5 \cdot z_t^{(k)}})$
  Composite: $S_{\text{adaptive}} = \sum_k w_k \cdot s_t^{(k)}$
  Blend: $S_t = \text{confidence} \cdot S_{\text{adaptive}} + (1 - \text{confidence}) \cdot S_{\text{fixed}}$
end for

3.5 Adaptive Baseline Algorithm

Our adaptive algorithm has three phases:

Key design choices:

- **Welford's algorithm** (Welford, 1962): Numerically stable online computation of running mean and variance. No batch statistics required.
- **Calibration window of 10 voiced frames**: Approximately 10–20 seconds of speech. Short enough for practical use; long enough for stable variance estimates.
- **EMA tracking ($\alpha = 0.02$)**: Slow adaptation to gradual changes in speaking style while being robust to transient stress spikes.
- **Confidence blending**: Smooth transition from fixed thresholds (during calibration) to adaptive scoring (post-calibration). Prevents discontinuities.
- **Sigmoid transform (steepness 1.5)**: Maps z-scores to $[0, 1]$ with center at $z = 0$ (the speaker's own mean). Steepness of 1.5 ensures that 1σ deviations produce moderate stress scores ($\sim$0.82).

3.6 Age-Adjusted Normalization

For pediatric applications, we define age-group-specific parameters:

Table 1: Age-adjusted reference ranges for voice stress features.

Parameter	Child 3–5	Child 5–10	Child 10–14	Adult
F0 range (Hz)	250–400	200–350	180–300	75–300
Normal HR (bpm)	80–120	70–110	60–100	60–100
Elevated HR (bpm)	≥ 130	≥ 120	≥ 110	≥ 100
Normal HRV SDNN (ms)	$\sim$65	$\sim$75	$\sim$85	$\sim$100

These ranges are used during the calibration phase to set initial priors and to filter out F0 estimates outside the physiologically plausible range for the speaker's age group.

4 Results

4.1 Pitch Detection Comparison

Table 2: F0 detection rate on SEP-28K (20,131 clips).

Method	Clips with valid F0	Detection Rate
Naive autocorrelation	2,437	12.1%
YIN (our implementation)	14,645	72.7%
YIN on clips with speech	14,645 / 14,941	**98.1%**

The naive autocorrelation method detects F0 in only 12.1% of clips. This is due to sub-harmonic errors, noise sensitivity, and lack of the CMNDF normalization that YIN provides. When restricted to clips containing detectable speech (excluding silence and noise-only clips), YIN achieves 98.1% detection.

4.2 Fixed vs. Adaptive Stress Distributions

Table 3 compares the stress score distributions produced by fixed and adaptive approaches.

Table 3: Stress score distribution statistics ($N = 14{,}645$).

Method	Mean	Std Dev	$\% \geq 0.8$	$\% \leq 0.2$
Fixed thresholds	0.731	0.187	61.4%	2.6%
Adaptive baseline	0.530	0.162	6.2%	0.1%

The fixed-threshold approach scores **61.4% of clips as high-stress** (≥ 0.8). Given that these are podcast speakers in a familiar, comfortable recording context, this proportion is implausibly high and likely reflects systematic overestimation due to inter-speaker vocal variability rather than genuine stress. The distribution is right-skewed with mean 0.731.

The adaptive approach produces a near-symmetric distribution centered at 0.530 with standard deviation 0.162. Only 6.2% of clips are scored as high-stress, and the distribution is approximately symmetric—more consistent with what one would expect from speakers in a low-stress recording context.

4.3 Within-Speaker Dynamics

To verify that the adaptive baseline preserves genuine stress variations, we examined stress score trajectories within individual podcast episodes. Within-speaker standard deviation ranges from 0.08 to 0.22 (mean 0.14), indicating that the adaptive approach retains sensitivity to within-speaker state changes while removing between-speaker baseline differences.

4.4 Calibration Convergence

Welford's algorithm reaches stable variance estimates ($< 5\%$ relative change) after 7–8 voiced frames on average, well within our 10-frame calibration window. The confidence blending ensures smooth transition from fixed to adaptive scoring.

5 Discussion

5.1 Sources of Skew in Fixed-Threshold Scoring

Fixed thresholds produce skewed distributions because they assume a homogeneous population with consistent vocal characteristics. In practice:

1. **Inter-speaker variability**: Baseline jitter ranges from $< 0.5\%$ to $> 5\%$ across speakers. A speaker with naturally high jitter ($\sim$4%) would be permanently scored as "stressed" under fixed thresholds.

2. **Recording conditions**: Podcast audio quality varies (professional studio vs. home recording), affecting measured shimmer and noise floor.

3. **Age and gender**: Children's voices have systematically higher F0, higher jitter, and greater F0 variability than adults. Fixed adult-derived thresholds are inappropriate for pediatric use.

5.2 Implications for Pediatric Applications

The case for adaptive baselines is strongest in pediatric speech monitoring:

- A 5-year-old child's resting F0 ($\sim$300 Hz) would produce "elevated" F0 variability under adult-calibrated fixed thresholds.
- Children's vocal motor control develops throughout childhood; what constitutes "normal" jitter varies by age.
- The adaptive baseline automatically accounts for these differences without requiring explicit age-specific threshold engineering.

Age-adjusted F0 ranges (Table 1) serve as guard rails for the calibration phase, filtering out physiologically implausible F0 estimates.

5.3 Design Choices

Why Welford over batch statistics? In a real-time monitoring application, the system must begin providing stress scores within seconds of starting a recording. Welford's algorithm provides stable running statistics without requiring a buffer of historical data.

Why EMA for tracking? After calibration, the speaker's baseline may drift (e.g., fatigue, warming up). EMA with $\alpha = 0.02$ provides slow adaptation (half-life $\approx$ 34 frames $\approx$ 1–2 minutes of speech), which tracks gradual changes without being destabilized by individual high-stress utterances.

Why sigmoid with steepness 1.5? The sigmoid maps z-scores to $[0, 1]$. Steepness of 1.5 ensures that the useful range of the output is not compressed. A steepness of 1.0 would be too gradual (90% of values between 0.27 and 0.73); 2.0 would be too aggressive (creating near-binary outputs). 1.5 provides a practical balance.

5.4 Limitations

1. **Cold start**: During the first 10–20 seconds, the system relies on fixed thresholds blended with incomplete adaptive estimates. This is mitigated by the confidence blending but not eliminated.

2. **Single-speaker assumption**: The algorithm assumes one speaker per session. Multi-speaker environments would require speaker diarization.

3. **No ground truth for stress**: SEP-28K does not include stress annotations. We evaluate distribution quality rather than classification accuracy.

4. **Adult data only**: The pediatric age-adjusted ranges are based on published norms, not validated on a pediatric disfluency dataset.

6 Conclusion

We showed that fixed thresholds for voice stress assessment produce implausibly skewed score distributions when applied across diverse speakers (61.4% scored as high-stress in an informal podcast corpus). Our adaptive baseline algorithm using Welford's online algorithm produces a near-symmetric stress score distribution ($\mu = 0.530$, $\sigma = 0.162$), consistent with more plausible calibration under heterogeneous speakers, while preserving within-speaker dynamics. Combined with YIN pitch detection (98.1% F0 rate vs. 12.1% for naive methods), this provides a practical foundation for voice stress scoring in real-time speech monitoring applications, with particular relevance for pediatric populations where fixed thresholds are most inappropriate.

The algorithm is implemented in DisfluoSDK, an on-device framework for speech disfluency detection and voice stress analysis on iOS.

Data Availability

The SEP-28K dataset is publicly available (Lea et al., 2021). Analysis code and the adaptive baseline implementation will be released at `https://github.com/AnonAuthor/adaptive-stress-baseline` 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, algorithm design, implementation, and analysis were conducted by the author.

References

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.

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.

Lee, L. & Rose, R. (1996). Speaker normalization using efficient frequency warping procedures. In *Proc. ICASSP*, 353–356. IEEE.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3), 1455–1468.

Nicollas, R., Garrel, R., Ouaknine, M., Giovanni, A., Nazarian, B., & Triglia, J. M. (2008). Normal voice in children between 6 and 12 years of age: Database and nonlinear analysis. *Journal of Voice*, 22(6), 671–675.

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

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

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