On-Device Disfluency Detection with Voice Stress Association Analysis: A Mobile Framework Validated on SEP-28K

Nazar Kozak

***Abstract*—We present DisfluoSDK, an on-device framework for real-time speech disfluency detection and voice stress analysis, motivated by privacy-sensitive speech therapy applications. The system classifies five disfluency types—prolongation, block, sound repetition, word repetition, and interjection—from 3-second mel-spectrogram clips using CoreML on iOS. We evaluate two architectures on the SEP-28K dataset (20,131 clips, 5-fold episode-grouped cross-validation): a custom 4-block convolutional neural network (617K parameters, 1.2 MB) achieving macro-F1 of 0.382, and an adapted ResNet-18 (11.2M parameters, 21 MB) achieving macro-F1 of 0.404. Both models achieve sub-millisecond inference (0.62–0.79 ms) on an Apple M1 processor, enabling 3,000-times real-time processing. We additionally examine the relationship between voice stress markers (jitter, shimmer, fundamental frequency variability) and disfluency labels across a subset of 14,645 clips with valid pitch estimates, finding no practically meaningful linear associations (all absolute correlations below 0.05, negligible effect sizes). To our knowledge, this is the first framework to perform multi-type disfluency classification entirely on-device.**

***Index Terms*—Disfluency detection, stuttering, on-device inference, CoreML, voice stress analysis, SEP-28K, mobile speech processing.**

I. Introduction

Speech disfluency—encompassing prolongations, blocks, repetitions, and interjections—affects approximately 5% of children during developmental years, with about 1% continuing to stutter into adulthood [1]. Early intervention significantly improves outcomes, yet access to speech-language pathologists (SLPs) remains limited, particularly in rural areas where the SLP-to-population ratio can exceed 1:100,000 [2].

Automated disfluency detection has the potential to democratize speech therapy by providing continuous, objective monitoring outside clinical settings. However, existing approaches predominantly rely on server-based processing, which introduces latency, requires reliable internet connectivity, and raises privacy concerns—particularly when processing voice data from vulnerable populations [3].

We present DisfluoSDK, an on-device framework for disfluency detection and voice stress analysis on iOS, motivated by future pediatric speech therapy applications. While our current evaluation uses adult speech data (SEP-28K), the architecture is designed with pediatric deployment as a target. Our contributions are threefold:

N. Kozak is with Kozak Technologies Inc, Los Angeles, CA 90036, USA. E-mail: nzrkzk@gmail.com

1) **On-device disfluency detection**: We train and deploy CNN models that classify five disfluency types from 3-second mel-spectrograms in under 1 ms on consumer hardware, with no server dependency.
2) **Adaptive voice stress analysis**: We implement a YIN-based pitch detector with per-speaker adaptive baseline calibration using Welford's online algorithm, enabling voice stress assessment that adapts to individual vocal characteristics.
3) **Empirical association analysis**: Across 14,645 clips with valid F0, we find no practically meaningful linear associations between voice stress markers and disfluency labels (all $|r| < 0.05$, negligible effect sizes), supporting the architectural separation of anxiety assessment from disfluency detection.

II. Related Work

A. Stuttering Event Detection

The SEP-28K dataset [4] established a benchmark for stuttering event detection from podcast audio, providing 28,177 annotated 3-second clips across five disfluency types. Lea et al. [4] reported per-type F1 scores using a ConvLSTM with mel-filterbank features augmented with F0 and articulatory features: Prolongation 0.685, Block 0.559, Sound Repetition 0.632, Word Repetition 0.604, Interjection 0.713. These results used a fixed random train/test split.

Bayerl et al. [5] demonstrated that 4 podcast hosts account for 59% of SEP-28K clips, and that random splits allow speaker overlap that inflates results. Speaker-independent evaluation substantially reduces detection performance. In subsequent work, Bayerl et al. [6] explored wav2vec 2.0 features with multi-task learning on an extended SEP-28K-E variant, achieving higher scores but requiring 300M+ parameter models unsuitable for mobile deployment. Sheikh et al. [7] provided a comprehensive review of machine learning approaches for stuttering identification, surveying detection architectures, feature representations, and evaluation methodologies.

B. On-Device Speech Processing

On-device speech processing has advanced with mobile inference engines such as CoreML [8], TensorFlow Lite [9], and ONNX Runtime [10]. While ASR has been deployed on-device [11], disfluency-specific detection remains predominantly server-based. To our knowledge, DisfluoSDK is the first framework to perform multi-type disfluency classification entirely on-device.

TABLE I
DATASET STATISTICS AFTER QUALITY FILTERING.

Disfluency Type	Positive	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%

C. Voice Stress Analysis

Voice stress indicators—jitter, shimmer, and F0 variability—have been studied as correlates of psychological stress [12], [13]. In stuttering populations, the relationship between voice stress markers and disfluency events has not been systematically examined at scale, leading to potential confounds in multimodal assessment systems.

III. DATASET

We use SEP-28K [4], containing 28,177 three-second clips from podcasts featuring people who stutter. Each clip is annotated by three raters for five disfluency types: Prolongation, Block, Sound Repetition, Word Repetition, and Interjection.

After filtering clips flagged for poor audio quality, music, or absence of speech (majority vote $\geq$ 2/3), and matching with available audio (258 of 385 episodes downloaded), we retain **20,131 clips** from five shows. Binary labels per disfluency type are established by majority voting ($\geq$ 2/3 annotators).

Each 16 kHz mono WAV clip is converted to a log-mel spectrogram with 128 mel bands, 1024-sample FFT, and 512-sample hop, yielding tensors of shape $(1, 128, 94)$. Per-sample normalization (zero mean, unit variance) is applied at training time.

IV. METHOD

A. Model Architectures

We evaluate two architectures for multi-label disfluency classification:

DisfluencyCNN. A 4-block convolutional network. Each block: two 3×3 convolutions with batch normalization, ReLU, 2×2 max pooling, and spatial dropout. Channels: $32 \to 64 \to 128 \to 256$, followed by adaptive average pooling and a classifier ($256 \to 128 \to 5$). Parameters: 616,549.

ResNet-18 Adapted. A ResNet-18 [14] pretrained on ImageNet, with the input convolution modified from 3 to 1 channel (averaging pretrained weights) and the head replaced with $512 \to 128 \to 5$. Parameters: 11,236,549.

Both output 5 logits for multi-label prediction via sigmoid activation and per-class thresholding.

B. Training

We use binary cross-entropy loss with per-class positive weights (ratio of negative to positive samples). AdamW optimizer with weight decay 10^{-4} and cosine annealing schedule. Training for up to 30 epochs with early stopping (patience = 7).

Data augmentation follows SpecAugment [15]: time masking (12 frames $\approx$ 0.4 s), frequency masking (20 mel bands), and random gain ($\pm$20%).

C. Episode-Grouped Cross-Validation

We employ 5-fold stratified group cross-validation grouped by podcast episode. This ensures all clips from the same episode (and thus the same recording session) remain in the same fold, reducing—though not fully eliminating—data leakage through shared speaker characteristics, as the same speaker may appear across multiple episodes [5].

D. Threshold Optimization

Rather than a fixed 0.5 threshold, we optimize per-class thresholds to maximize F1 on the validation set—important given class imbalance shifts the operating point.

E. On-Device Deployment

Trained models are exported to CoreML (.mlpackage) via TorchScript tracing and coremltools, targeting iOS 17+. Per-class thresholds are embedded as model metadata.

V. VOICE STRESS ANALYSIS

A. YIN Pitch Detection

F0 extraction uses the YIN algorithm [16], computing a cumulative mean normalized difference function (CMNDF). Our implementation uses Apple's Accelerate framework (vDSP): precompute cumulative energy, compute autocorrelation via `vDSP_dotpr`, calculate the difference function $d(\tau) = E_{\text{head}} + E_{\text{tail}} - 2 \cdot \text{autocorr}(\tau)$, and apply CMNDF with threshold $\tau_{\text{th}} = 0.15$. Frame size: 1024 samples (64 ms at 16 kHz), hop: 512 samples. This achieves 98.1% F0 detection across SEP-28K, versus 12.1% with naive autocorrelation.

B. Adaptive Baseline

We implement per-speaker adaptive baseline calibration:

- **Calibration** (first 10 voiced frames): Welford's online algorithm computes running mean and variance for jitter, shimmer, F0 std, and speech rate.
- **Tracking**: Exponential moving average ($\alpha = 0.02$) for drift adaptation.
- **Scoring**: Z-score $\to$ sigmoid (steepness = 1.5) $\to$ weighted composite, blended with raw scores proportional to calibration confidence.

C. Independence from Disfluency

We computed Pearson and point-biserial correlations between all voice stress features and disfluency labels across $N = 14{,}645$ clips with valid F0 (Table II).

While four Prolongation correlations reach statistical significance due to the large N, all effect sizes are **negligible** by Cohen's convention ($|r| < 0.10$). The strongest (stress $\times$ prolongation, $r = -0.050$) explains only 0.25% of variance ($r^2 = 0.0025$). We did not observe practically meaningful

TABLE II
PEARSON CORRELATION (r) BETWEEN VOICE STRESS FEATURES AND DISFLUENCY LABELS ($N = 14{,}645$). *SIGNIFICANT AFTER BONFERRONI CORRECTION ($\alpha = 0.05/20$).

Feature	Pro	Blk	SRep	WRep	Intj
Jitter	$-.028^*$	$-.006$	.008	.013	.001
Shimmer	$-.030^*$	.002	.019	.013	$-.002$
F0 Std	$-.028^*$	$-.007$	$-.009$	.021	.001
Stress	$-.050^*$	$-.013$	$-.007$	.025	.004

TABLE III
PER-TYPE RESULTS, 5-FOLD CV, THRESHOLD = 0.5.

	CNN				ResNet-18			
Type	P	R	F1	AUC	P	R	F1	AUC
Pro	.276	.737	.401	.841	.266	.718	.388	.829
Blk	.201	.555	.295	.671	.196	.531	.287	.658
SRep	.222	.630	.328	.777	.233	.648	.342	.790
WRep	.174	.643	.273	.669	.190	.686	.298	.717
Intj	.411	.638	.500	.738	.461	.674	.548	.792
Macro	**.257**	**.640**	**.360**	—	**.269**	**.651**	**.373**	—

TABLE IV
RESULTS WITH OPTIMIZED PER-CLASS THRESHOLDS.

	CNN		ResNet-18	
Type	F1	Thresh	F1	Thresh
Pro	.461	0.69	.453	0.70
Blk	.299	0.53	.291	0.55
SRep	.368	0.67	.385	0.71
WRep	.279	0.55	.324	0.66
Intj	.505	0.52	.564	0.63
Macro	**.382**	—	**.404**	—

TABLE V
COMPARISON ON SEP-28K (F1 PER TYPE). EVALUATION PROTOCOLS DIFFER.

Method	Protocol	Pro	Blk	SR	WR	Intj	Params	Device
Lea [4]	Rand.	.685	.559	.632	.604	.713	$\sim$2M	No
SVM [5]	Spkr-ind.	.460	.360	.460	.510	.710	—	No
Bayerl [6]	Spkr-ind.	.530	.320	.530	.640	.770	300M+	No
CNN	Ep-grp	.461	.299	.368	.279	.505	**617K**	**Yes**
RN18	Ep-grp	.453	.291	.385	.324	.564	11.2M	**Yes**

linear associations between voice stress features and disfluency labels in this dataset. This supports—but does not conclusively prove—the architectural decision to treat stress analysis and disfluency detection as separate modules, as non-linear or conditional relationships cannot be ruled out from marginal correlations alone.

VI. EXPERIMENTS AND RESULTS

A. Disfluency Detection

Table III shows per-type results at default threshold 0.5, and Table IV with optimized thresholds.

Threshold optimization improves macro F1 by 6.1% (CNN) and 8.3% (ResNet-18). Per-fold stability: CNN 0.359 ± 0.011, ResNet-18 0.377 ± 0.014.

B. Comparison with Prior Work

Lea et al.'s higher F1 scores are partly attributable to their random split, which permits speaker overlap [5]. Our episode-grouped protocol is more conservative than random splitting but less strict than full speaker-independent evaluation (the same speaker may appear in multiple episodes). Compared to the speaker-independent SVM baseline [5], our ResNet-18 achieves comparable Prolongation (0.453 vs. 0.460) while being the first to operate on-device with sub-millisecond inference.

C. On-Device Performance

Both models achieve sub-millisecond CoreML inference, with the CNN offering a 1.2 MB footprint suitable for bandwidth-constrained deployment. Real-time factors exceed 3,700×, leaving ample compute for VAD, stress analysis, and UI rendering.

VII. DISCUSSION

A. Performance Context

Several factors contribute to moderate F1 scores: (1) low inter-annotator agreement (Fleiss' κ: prolongation 0.11, block 0.25 [4]), establishing an empirical ceiling; (2) variable podcast audio quality; (3) some events span <0.5 s within 3 s clips. Notably, our per-type performance ranking mirrors inter-annotator agreement, suggesting detection difficulty is intrinsic to the types.

The speaker-independent SVM [5] achieves macro F1 $\approx$ 0.50, closer to our 0.40 than to Lea et al.'s 0.64. The remaining gap reflects an accuracy–efficiency trade-off: Lea et al. use temporal modeling, augmented features, and CCC loss. Our models sacrifice these for a 1.2 MB, sub-ms footprint.

B. Clinical Implications

The absence of meaningful linear associations between voice stress and disfluency (Section V) suggests that these signals can be treated as complementary rather than redundant in multimodal systems. If this finding generalizes to pediatric populations—which remains to be validated—it would support architectures that maintain separate modules for disfluency tracking and communication anxiety monitoring.

C. Privacy

All inference occurs locally on the user's device. No audio data is transmitted, which reduces privacy risk—a particularly relevant consideration for future pediatric deployments where children's voice data is involved. On-device processing also provides sub-ms latency (versus 100+ ms cloud round-trips) and availability without network connectivity. We note that on-device inference alone does not constitute full regulatory compliance (e.g., COPPA), which additionally requires appropriate consent flows, data storage policies, and telemetry controls.

TABLE VI
INFERENCE BENCHMARKS (COREML, APPLE M1, MACOS 15).

Model	Size	CoreML	Speedup	RT Factor
CNN	1.2 MB	0.62 ms	9.1×	4,839×
ResNet-18	21 MB	0.79 ms	9.6×	3,797×

D. Limitations

(1) We evaluate on 71.4% of SEP-28K (missing episode downloads). (2) SEP-28K contains adult speakers; pediatric validation (e.g., FluencyBank) is needed. (3) Optimized thresholds may require recalibration for different populations. (4) No temporal modeling across consecutive clips.

VIII. CONCLUSION

We presented DisfluoSDK, achieving macro F1 of 0.404 on SEP-28K with sub-ms CoreML inference on Apple M1. Key contributions: (1) on-device disfluency classification in 1.2–21 MB, (2) adaptive voice stress analysis with per-speaker calibration, and (3) empirical evidence that voice stress markers and disfluency labels lack practically meaningful linear associations in adult speech. Future work includes validation on pediatric speech (FluencyBank), lightweight temporal modeling to improve detection, and clinical pilot studies with speech-language pathologists.

Code and trained models are available from the author upon reasonable request.

ACKNOWLEDGMENTS

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

REFERENCES

[1] E. Yairi and N. Ambrose, "Epidemiology of stuttering: 21st century advances," *Journal of Fluency Disorders*, vol. 38, no. 2, pp. 66–87, 2013.
[2] K. Wylie, L. McAllister, B. Davidson, and J. Marshall, "Changing practice: Implications of the World Report on Disability for responding to communication disability in under-served populations," *International Journal of Speech-Language Pathology*, vol. 15, no. 1, pp. 1–13, 2013.
[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.
[4] C. Lea, V. Mitra, A. Joshi, S. Kajarekar, and J. P. Bigham, "SEP-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6798–6802.
[5] S. P. Bayerl *et al.*, "The influence of dataset partitioning on dysfluency detection systems," in *Proc. ACL Workshop on Computational Approaches to Linguistic Code-Switching*, 2022.
[6] ——, "Cross-corpus stuttering detection as a multi-label problem," in *Proc. Interspeech*. ISCA, 2023.
[7] S. A. Sheikh *et al.*, "Machine learning for stuttering identification: Review, challenges, and future directions," *Neurocomputing*, vol. 514, pp. 385–402, 2022.
[8] Apple Inc., "Core ML framework documentation," https://developer.apple.com/documentation/coreml, 2023.
[9] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, 2016, pp. 265–283.
[10] ONNX Runtime, "ONNX Runtime Mobile," https://onnxruntime.ai, 2023.
[11] Y. He *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.
[12] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.
[13] K. Pisanski *et al.*, "Voice parameters predict sex-specific body morphology in men and women," *Animal Behaviour*, vol. 112, pp. 13–22, 2016.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*. ISCA, 2019, pp. 2613–2617.
[16] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.