

Contrastive Speech Analytics & Temporal Flaw Grounding: Technical Documentation

Multimodal AI Hackathon 2026, Track C. Built on the AI Dataset Maker alignment pipeline.

1. Problem and approach

Judging spoken performance is subjective: two judges hear the same speech and disagree, and the speaker gets feedback like "work on your pacing". We make delivery feedback measurable, time-stamped and explainable:

Compare a reading against a reference delivery of the same text, word by word, on speaker-normalized acoustic scales, and report every region that deviates, with its start, end and the numbers behind it.

The system has four parts: a contrastive dataset with exact labels (§2), forced alignment + feature extraction (§3), temporal grounding + causal explanation (§4), and a

reproducible rubric (§5). We validate the detector on the dataset itself (§6) and serve it in an interactive dashboard (§7).

2. Contrastive dataset

2.1 "Good" baseline

Three public-domain speeches by celebrated speakers (U.S. federal works, sourced from Wikimedia Commons), each excerpted to 32–57 s:

id	Speaker	Speech	Words	Align score
fdr_infamy	Franklin D. Roosevelt	"Day of Infamy", 1941 (opening)	82	0.79
reagan_challenger	Ronald Reagan	Challenger address, 1986 (opening)	107	0.92
jfk_moon	John F. Kennedy	Rice University, 1962 (closing)	83	0.76

Three speakers, three decades and three recording chains test speaker agnosticism. Transcripts are the official texts, not ASR output: Whisper hallucinated on these old recordings (repeated "Bye." over applause), which is exactly the kind of noise this dataset must avoid.

2.2 "Bad" spectrum: controlled flaw injection

Every flawed file is produced from its baseline by a signal operation whose position is chosen deterministically (hash of speech id + flaw type), so the same words carry the flaw at every severity, giving a clean gradient on identical content:

Flaw	Operation	Severity 1 → 2 → 3
rushed	PSOLA time compression (pitch preserved), 7 words	duration ×0.80 → 0.65 → 0.50
dragging	PSOLA time stretch (pitch preserved), 7 words	×1.30 → 1.60 → 2.00
long_pause	room-tone silence inserted between two words	+0.8 → 1.6 → 2.8 s
monotone	Praat Manipulation: F0 pulled toward its mean	60% → 85% → 100% flatter
mumble	gain + Hann band-pass (quieter, duller)	−8 dB/3 kHz → −14/1.8 → −20/1.0
stutter	word onset (≥120 ms) repeated with room-tone gaps	1 → 2 → 3 repeats

Plus two composites per speech: near_perfect (two severity-1 flaws) and botched (four severity-3 flaws at once).

Total: $3 \times (1 + 18 + 2) = 63$ files, 39 MB FLAC.

Exact labels. Each operation keeps a piecewise-linear time map from the original timeline to the new one (crossfades at every splice are accounted for), so every flaw's start/end and every word's timestamp in the flawed file are known exactly, without re-annotation. Inserted silence uses the recording's own room tone (its quietest 200 ms, tiled), so pauses are not trivially detectable as digital zeros.

Files. `contrast_data/dataset/<speech>/<name>.flac` and `<name>.json` (transcript, source, license, word timestamps, flaw regions), with `metadata.csv` indexing all files. Rebuild: `python -m contrast.build_dataset`.

3. Forced alignment and features

Alignment. Both recordings are aligned to the same transcript with Meta's MMS_FA (torchaudio), using a wildcard token `*` between every pair of words. The wildcard absorbs sound that is not in the text (fillers, repeated onsets), which is what we use to detect disfluency. Numbers are spelled out (num2words) because the aligner's vocabulary is `a-z`.

Frame features (10 ms grid).

Feature	Tool	Normalization (speaker agnostic)
F0	Praat autocorrelation pitch (75–500 Hz)	semitones re the speaker's median F0
Energy	Praat intensity	dB re the speaker's median speech energy
HNR (voice clarity)	Praat cross-correlation harmonicity	dB
Spectral centroid	librosa (FFT, 512)	Hz, used as a ratio vs baseline
MFCC $\times 13$	librosa	cepstral mean normalized

Word features. Articulation rate (syllables / speaking time over a 3-word window; silences longer than 0.2 s are excluded), pause before the word, pitch movement (SD of semitones over a 5-word window), mean energy, centroid, HNR, and the word's MFCC sequence.

4. Temporal grounding and causal explanation

One delta table. Word k of the participant corresponds to word k of the baseline, so we build a single per-word table of deltas: rate log-ratio, extra pause, pitch-movement ratio, energy dB, MFCC-DTW articulation distance, and extra sound before the word. Every later step reads this table.

Style profile first. The medians of the deltas describe the participant's global style: overall tempo, pause length and pitch range relative to the reference. Local flaws are measured

against this profile: a speaker who is 15% slower everywhere gets one finding ("overall slow tempo"), not 80 flagged words. This is what makes the comparison fair across voices.

Local deviations. For each feature we compute the deviation in natural units and a z-score against the measurement noise of this pair: the robust SD ($MAD \times 1.4826$) of the word-by-word differences. Most words are delivered the same way, so this scale is what "normal disagreement" looks like.

A word is flagged only when both an effect-size threshold and $|z| > 1.5$ pass:

Flaw	Effect-size rule
rushed / dragging	rate ratio > 1.18 or $< 1/1.18$
long_pause	extra silence > 0.45 s
monotone	pitch-movement ratio < 0.62 (where baseline moves > 1 st)
mumble	energy < -5 dB vs baseline (centroid drop reported as supporting evidence)
stutter	voiced wildcard sound before a word exceeding the baseline's by > 0.15 s
unclear	(a) the aligner fits $< 50\%$ of the word's reference duration and the rest is voiced sound that does not match it, or (b) per-word normalized MFCC-DTW distance ($c1-c12$) $> 1.8 \times$ the pair median and $z > 5$

Consecutive flagged words merge into one region (start of the first word to the end of the last). A clean-up pass enforces one cause, one region: a stutter suppresses the "slow" rate it causes on its word, and the gaps inside a stutter or a dragged phrase are not reported again as long pauses. Articulation ("unclear") is reported only when no other flaw explains the spectral change: muffling and time-scaling also move MFCCs. Minimum durations remove word-boundary artefacts.

Every region carries a causal explanation built from its numbers, for example:

12.4–15.1 s · Rushed pace. Speech rate 8.3 syl/s vs baseline 5.5 syl/s (+52%, $z=+8.5$).
Words are compressed; listeners lose the content.

5. Rubric and coaching (reproducible)

Five criteria, 0–10: Pace, Pausing, Expressiveness, Volume & clarity, Fluency. Each region adds a penalty magnitude $\times (0.5 + 0.5 \times \min(1, \text{duration} / 5 \text{ s}))$, where magnitude is the normalized deviation (for example $|\log \text{rate ratio}| / \log 2$, extra pause / 2.5 s, $1 - \text{pitch-movement ratio}$). Whole-speech habits from the profile add their own penalty (tempo outside $\pm 15\%$, pauses over $1.4\times$ the reference, pitch range under 80%). Criterion = $10 \times \max(0, 1 - \Sigma \text{penalty})$, overall = mean. No sampling and no LLM: the same input gives the same score on every run.

The rubric returns every penalty as an item with its source (a region or a profile habit). The coach uses that list directly:

- Tip per flaw, built from the region's own numbers and target, for example: `"Slow down on 'the state of the Union, but': aim for about 6.3 syllables/s (you were at 9.2)", or "lift your pitch on 'This' (the word the reference stresses most in this phrase) by about 1 semitone"`.
- Practice plan: the three items that cost the most points, in order. What to practise first therefore follows the score, not a separate heuristic.
- A/B clips: for any region, the same word span [i..j] is cut from both recordings, so you hear the reference phrase and then yours.
- PDF report: rubric, plan, profile, word heatmap and every flaw with its explanation and tip, rendered from the same objects as the dashboard.

```
compare() -> profile --> regions [i..j] + numbers --> word heatmap
|
+-----> rubric (itemized) <-+
|
tips . practice plan . A/B clips --> PDF report
```

6. Stress test: does it find what we injected?

`python -m contrast.stress_test` aligns every file from scratch (the detector never sees the labels) and scores predictions against the ground truth. Full output: `contrast_data/stress_report.md`.

63 files · recall 81% (severity 1: 71%, 2: 89%, 3: 83%) · precision 66% · false alarms on clean baselines: 0

Flaw	sev 1	sev 2	sev 3	start err (median)	end err (median)	score s1 → s2 → s3
dragging	3/3	3/3	2/3	0.078 s	0.273 s	9.47 → 9.03 → 8.8
long_pause	3/3	3/3	5/6	0.032 s	0.027 s	9.37 → 9.13 → 8.37
monotone	5/6	3/3	6/6	0.335 s	0.383 s	9.5 → 8.97 → 8.73
mumble	3/3	3/3	3/3	0.119 s	0.021 s	8.93 → 8.27 → 7.17
rushed	3/6	3/3	6/6	0.051 s	0.02 s	9.77 → 9.33 → 9.1
stutter	0/3	1/3	3/6	0.045 s	0.226 s	9.87 → 9.9 → 9.6

Rubric falls monotonically with severity for 5/6 flaw types. Composite files: `{'baseline': 10.0, 'near_perfect': 9.0, 'botched': 5.33}`

Reading the table. Rows are flaw types and columns are severities (counts are found/injected). A flaw is "found" when a predicted region of the same type overlaps its injected timestamps ($\text{IoU} \geq 0.2$; any overlap for point-like pauses and stutters). Boundary errors are medians over found flaws. The last column is the mean overall rubric score

at each severity.

What works. Zero false alarms on the three clean baselines. Pauses, pace and volume are localized within tens of milliseconds. Severity 2–3 flaws are found 83–89% of the time. The rubric falls with severity for 5 of 6 flaw types, and from 10.0 (baseline) to 9.0 (near-perfect) to 5.3 (botched).

What doesn't (yet). Stutter is the weak class (4/12): a single repeated onset is short, and the aligner often folds it into the neighbouring word, where it reads as slow pace. Subtle severity-1 rushed readings ($\times 0.8$) are found 3/6 times. Precision (66%) is lowered mostly by extra regions in the botched composites, where four overlapping flaws shift the alignment.

Articulation probe. The synthetic set contains no mispronunciations, so we reverse one content word per position (same energy and spectrum, wrong sound order; 3 per speech). 4/9 are reported as unclear articulation at the right word and 7/9 are flagged as some flaw at that spot; 2/9 are missed. This is the least mature detector.

7. Dashboard

python contrast_app.py (Gradio):

1. Pick a reference speech (or upload your own audio + transcript). 2. Record with the microphone or upload a reading of the same text, or pick any file from the dataset (its hidden ground truth is shown after analysis, for comparison). 3. See the rubric cards, the practice plan and your style profile (tempo, pauses, pitch range, voice). 4. Read the word heatmap: every word colored by its largest deviation; hover for the numbers. 5. Inspect the 3-panel time series (pitch, energy, speech rate) with the reference time-warped onto your timeline through the word alignment, so the curves line up word for word. Flaw regions are shaded and numbered. 6. In the flaw table (timestamp, numeric cause, what to do), click a row to hear A/B: the reference phrase, then yours. Download everything as a PDF report.

8. Limitations

- The bad spectrum is synthetic. It isolates each flaw cleanly, which is what makes the labels exact, but real flaws co-occur and interact. Self-recorded flawed readings are the natural next step.
- The reference must be a reading of the same text. Delivery is judged relative to one good performance, not against an absolute norm.
- Three speakers, all U.S. male orators from historical recordings. Wider diversity is needed to claim full speaker agnosticism.
- Syllable counts are a vowel-group estimate. That is fine for rate ratios, but not exact.
- Articulation ("unclear") is the least validated detector: the synthetic set has no mispronunciations, so the stress test adds a probe instead (one word time-reversed: same energy and spectrum, wrong sound order). Very short words (< 0.15 s) are not caught.
- Severity-1 flaws are subtle by design. Lower recall there is expected and reported, not tuned away.
- No held-out test set. Thresholds were tuned while inspecting files from this same dataset (mainly the Reagan and FDR excerpts), so the stress-test numbers are optimistic. The honest next step is a held-out speaker and self-recorded flawed readings.

9. Reproducibility

Pinned dependencies (requirements.txt), deterministic flaw placement, no random sampling in features or scoring, CPU-only operation (GPU optional).

```
pip install -r requirements.txt praat-parselmouth librosa plotly num2words gradio
python -m contrast.build_dataset # rebuild the 63-file dataset from raw/
python -m contrast.stress_test # metrics
python contrast_app.py # dashboard on :7861
```

Sources: [Wikimedia Commons](#) (public domain): [Roosevelt_Pearl_Harbor.ogg](#),
[Ronald_Reagan_speaks_on_the_Challenger_Space_Shuttle_Disaster.ogg](#),
[Jfk_rice_university_we_choose_to_go_to_the_moon.ogg](#).