Deepfake speech detection based on biological factors

Faculty of Information Technology, Brno University of Technology · 2026

Author: **Bc. Michal Frič**
Supervisor: **Ing. Vojtěch Staněk**

Excel@FIT 2026

Artefact-based detectors collapse on unseen synthesizers. **Do handcrafted physiological markers of human speech carry a usable, complementary signal?** We built 37 such markers and explored what actually works.

Motivation

State-of-the-art detectors chase *synthesis artefacts*; they spike in-domain but degrade when the synthesizer, codec, or channel changes. Human speech also carries *physiological* structure—respiration, stable phonation, timing—that is hard to fake consistently. **Hypothesis:** inconsistencies in these biological patterns can flag deepfakes and *complement* artefact-based systems. This poster summarises an **exploration**: whether 37 handcrafted markers carry usable signal, and which families matter.

Table 1 · EDS-14 (English, $\geq$ 14 s)

Dataset	Bona fide	Deepfake	Split
ASVspoof 5	9 208	20 121	train/val/test
Codecfake	0	79	train/val/test
In-the-Wild	194	421	train/val/test
SCDF	0	1 373	train/val/test
SpeechFake	1 947	46 932	**held-out OOD**
M-AILABS + MLAAD	849	1 159	**held-out OOD**
Total (EDS-14)	**12 198**	**70 085**	–

Held-out OOD (SpeechFake, M-AILABS+MLAAD): same leave-one-corpus rule — train/val/test are drawn only from the other EDS-14 sources; the classifier never trains on the held-out corpus.

Figure 1 · Pipeline

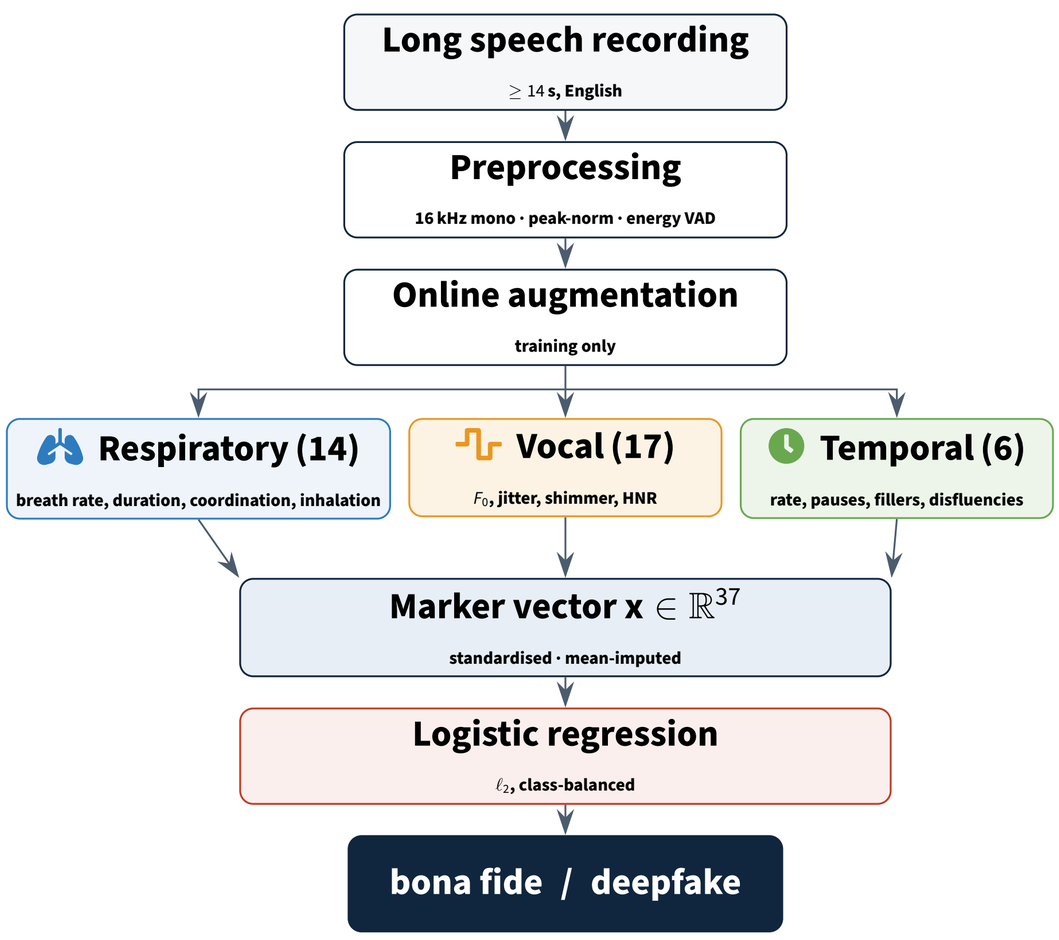
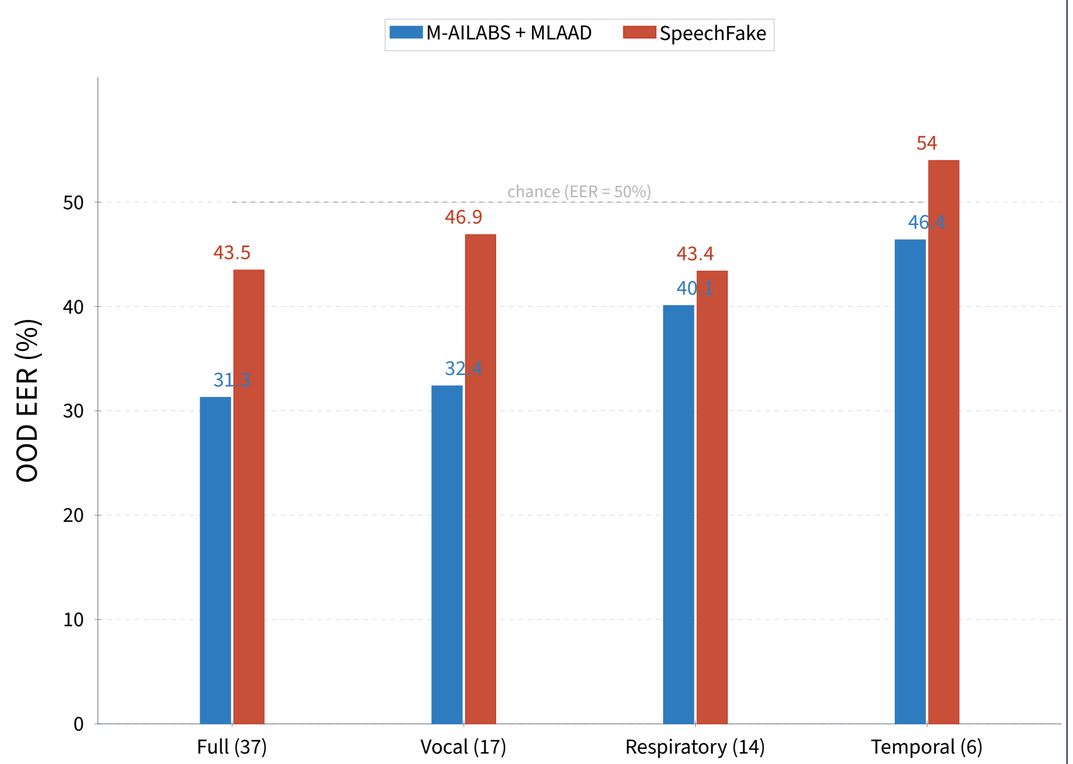


Chart 3 · Headline Results *(logreg, aug)*

In-domain (test split)	OOD · M-AILABS + MLAAD (*n*=2 008)	OOD · SpeechFake (*n*=48 879)
31.7%	**31.3%**	**43.5%**
EER	EER · AUC 0.75	EER · AUC 0.59

Take-away. The in-domain and MLAAD+M-AILABS EER are within one percentage point-the detector behaves almost the same on synthesizers it has seen and on ones it has not, while SpeechFake is markedly harder.

Chart 4 · Per-category Ablation *(OOD EER)*



Vocal carries nearly the full signal on M-AILABS+MLAAD. **Respiratory** is weaker but above chance. **Temporal** is **at or below chance** across OOD pools – the sub-stack needs to be redesigned.

Table 2 · 37 Physiological Markers

Respiratory (14)

breath rate, mean breath duration
pre-speech inhalation ratio
breath–speech coordination score
summary stats of breath / pause lengths

detector: Respiro-en

Vocal (17)

F_0 mean / median / std / range
jitter (local, RAP)
shimmer (local, APQ)
harmonics-to-noise ratio (HNR)

autocorrelation F_0 + pitch marks

Temporal (6)

speech rate, articulation rate
pause rate, mean pause duration
filler-word frequency
disfluency frequency

Whisper word-stamps + acoustic fallback

Experiments · What Was Swept

Not a single training run—a small **design sweep** on the EDS-14 pool (**Table 1**, top-right), then headline numbers for the chosen operating point (**Charts 3–4**).

Swept dimensions

- **Classifier:** logistic regression vs. rbf SVM (within ∼1 EER point).
- **Regularisation:** ℓ_2 strength C on a wide grid.
- **Augmentation:** μ-law codec simulation, RawBoost perturbations, and colored/Gaussian noise with random EQ-style filtering.
- **Imputation:** marker-wise mean vs. zero for failed extractions.
- **Feature groups:** all 37 vs. respiratory / vocal / temporal only.
- **OOD pools:** held-out M-AILABS+MLAAD and SpeechFake.
- **LDA:** on/off before the linear model (reported: **off**).

Reported setting. logreg, aug on, mean-imputation, no LDA.

Take-aways & Next Steps

- A simple **linear** classifier on 37 physiological markers reaches a non-trivial OOD score on unseen synthesizers-evidence that the cue is not a memorised artefact.
- Signal concentrates in **vocal** and, to a lesser extent, **respiratory** markers; temporal markers as currently implemented are at chance.
- **Next:** redesign the temporal sub-stack with a dedicated disfluency model; evaluate score-level fusion with an SSL back-end (e.g. WavLM) to quantify complementarity with artefact-based detectors.

Contact xfricm02@stud.fit.vutbr.cz · Computed on MetaCentrum / CESNET