BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Deepfake Speech Detection Based on Biological Factors

Michal Frič*

Abstract

State-of-the-art speech deepfake detectors rely on synthesis artefacts and their accuracy collapses once the synthesizer, codec, or recording condition changes; we ask whether a handcrafted, physiologically grounded representation can provide a complementary signal. We extract 37 *handcrafted physiological markers* (respiratory, vocal, temporal) from long English utterances, classify them with ℓ_2-regularised logistic regression, and explore the design space through classifier, C, imputation, augmentation, and per-category ablations. In-domain test EER is 31.7 %; on a strict out-of-distribution split of MLAAD and M-AILABS the detector reaches 31.3 % EER - essentially the same number - and the per-category ablation isolates the signal to *vocal* and, to a smaller extent, *respiratory* markers, while *temporal* statistics behave at chance. Handcrafted physiological markers are a usable and interpretable cue, worth keeping alongside artefact-based detectors, provided the temporal sub-stack is redesigned.

*xfricm02@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Modern text-to-speech (TTS) and voice-conversion (VC) systems make convincing speech deepfakes cheap to produce [1]. The dominant countermeasure is a data-driven classifier that learns to spot *synthesis artefacts* - spectral irregularities, phase discontinuities, vocoder fingerprints - in the acoustic signal [2]. These detectors are highly accurate in-domain but they lean on patterns tied to a specific synthesizer, codec, or recording condition, and therefore degrade on generalization [3].

Human speech, on the other hand, carries various *physiological markers* originating in the vocal apparatus itself: coordinated respiration, stable phonation, and temporal dynamics [4]. These markers are difficult to reproduce faithfully even in state-of-the-art synthesizers [5]. This motivates the core hypothesis of this work: *deepfake speech can be detected by identifying inconsistencies in the biological patterns of natural speech*. We therefore propose a detector that looks at how the voice was produced, not at how it was rendered, and is meant to *complement* current artefact-based systems.

This paper is an *exploration*: does such a handcrafted physiological representation carry any useful signal, and if yes, which markers actually contribute?

Contributions.

1. A deepfake speech detection approach based on 37 handcrafted physiological markers of human speech production, grouped into *respiratory*, *vocal* and *temporal* cues.
2. **EDS-14**, a reproducible English pooled corpus of utterances longer than 14s collected from six public sources under a strict dataset-level OOD protocol, together with the long-form feature-extraction pipeline.
3. An empirical exploration of the representation - classifier sweep, imputation strategy, augmentation on/off, per-category ablation - indicating that vocal and respiratory markers carry informative signal while temporal markers, as currently implemented, do not.

2. Physiological Markers

Why long recordings? Breathing, pause-rhythm coordination and disfluency statistics need *context*. We therefore only accept utterances longer than 14s. This is a price we pay on purpose: the short-utterance regime (4 s clips typical for ASVspoof) simply does not contain enough air cycles to measure breath coordination.

Three marker families. The 37 markers are grouped as follows:

Respiratory (14): breath rate, mean breath duration, pre-speech inhalation ratio, and a *coordination score* that quantifies how regular the alternation between breath offsets and speech onsets is. Breaths are localised with the Respiro-en neural detector [6]; all downstream statistics are deterministic.

Vocal (17): F_0 mean/median/standard deviation, jitter (local and RAP[1]), shimmer (local and APQ), and HNR [7]. These are computed from pitch marks produced by an autocorrelation-based F0 tracker.

Temporal (6): speech rate, articulation rate, pause rate, mean pause duration, and frequencies of *filler words* (*uh*, *um*, *hm*…) and *disfluencies* (repetitions, restarts). Fillers come from a Whisper word-timestamp pass with an acoustic fallback for low-confidence regions [8].

Preprocessing. Audio is resampled to 16 kHz mono, peak-normalised, and segmented into speech/silence by an RMS-energy VAD (2048-sample window, 512-sample hop, threshold 0.01).

Classifier. All 37 markers are concatenated into a single vector $\mathbf{x} \in \mathbb{R}^{37}$, standardised on the training set, and fed to ℓ_2-regularised logistic regression with class-balanced weights. The regularisation strength C is swept in Section 3; the reported setting is $C = 10^{-2}$. Missing markers (extraction failures) are replaced with zero value. Training waveforms are additionally passed through augmentation before marker extraction(μ-law codec simulation, RawBoost perturbations, and colored/Gaussian noise with random EQ-style filtering). **Figure 1**, **Table 2**.

3. Experiments

The goal is modest and exploratory: decide whether 37 handcrafted physiological markers carry any usable signal for deepfake detection, and, if they do, say *which* ones.

Dataset - EDS-14. We pool six public corpora and keep only English recordings longer than 14 s, yielding 12 198 bona-fide and 70 085 deepfake utterances. Sources span several generations of TTS and VC: ASVspoof 5 [1], Codecfake [9], In-the-Wild [3], the paired M-AILABS+MLAAD set [10, 11], SCDF [12], and SpeechFake [13]. **Table 1**.

Protocol. Two *mutually exclusive* strict-OOD protocols on the pooled English corpora: one hold-out is left out of the speaker-disjoint 70 / 15 / 15 train / val / test split and used only for OOD; **everything else** (including the other OOD corpus) trains in-domain. (i) M-AILABS+MLAAD held out ($n = 2\,008$)—SpeechFake is in the split; (ii) English SpeechFake slice held out ($n = 48\,879$)—M-AILABS+MLAAD is in the split. The two settings swap the excluded corpus, not one model with both held out.

What was swept. We toggled four design choices on the same merged vector: logistic regression vs. RBF SVM; waveform augmentation on/off before extraction; zero vs. train-split *median* imputation for failed markers; and all 37 features vs. one family at a time (*respiratory*, *vocal*, *temporal*). The reported pipeline is the best overall: full 37-dim logreg, augmentation on, median imputation, no LDA.

4. Results

With all 37 markers the detector reaches **31.7 % in-domain test EER**, **31.3 % EER on M-AILABS+MLAAD** (AUC 0.75), and **43.5 % EER on SpeechFake** (AUC 0.59). **Chart 3**. The near-zero gap on M-AILABS+MLAAD suggests the markers are not simply memorising synthesizer-specific artefacts.

Which markers help? Per-family retraining on M-AILABS+MLAAD OOD: *vocal*-only matches the full 37-dim model (32.4 % EER, AUC 0.74)-most separability lives in F_0/jitter/shimmer/HNR here. *Respiratory*-only is weaker (40.1 % EER, AUC 0.63) but clearly beats temporal-only. *Temporal*-only (pauses, disfluencies) is near random (46.4 % EER, AUC 0.54; worse on SpeechFake). **Chart 4**.

Role of augmentation. Turning augmentation off degrades OOD EER by 4-5 points while the in-domain EER moves by less than one point; robustness is paid for at training time, not by the classifier.

5. Conclusions

This work explored 37 handcrafted physiological markers for long-form English deepfake detection. The headline is *OOD-dependent*: the same logreg pipeline reaches ∼31 % EER on held-out M-AILABS+MLAAD but ∼44 % on an English SpeechFake OOD. Per-family retraining aligns with the main run-most separability, which sits in *vocal* features, less in *respiratory*; with only our current *temporal* summaries, the detector is effectively uninformative and should be reworked or omitted until better modelling exists. Proposed follow-ups are a stronger temporal/disfluency module and late fusion with an SSL artefact score to test complementarity.

Acknowledgements

I would like to thank my supervisor, Ing. Vojtěch Staněk, for his guidance and consultations.

[1] Relative Average Perturbation (RAP): a standard jitter measure quantifying short-term F_0 variability.

References

[1] Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8, 2024.

[2] Anton Firc, Kamil Malinka, and Petr Hanáček. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon*, 9(4):e15090, 2023.

[3] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Interspeech*, 2022.

[4] Nidula Elgiriyewithana and N. D. Kodikara. Attention-based efficient breath sound removal in studio audio recordings. In *Proceedings of the 5th International Conference on NLP & Information Retrieval*, pages 49–58, 2024.

[5] Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pages 4978–4982, 2024.

[6] Dong Yang, Tomoki Koriyama, and Yuki Saito. Frame-Wise Breath Detection with Self-Training: An Exploration of Enhancing Breath Naturalness in Text-to-Speech. In *Interspeech 2024*, pages 4928–4932, 2024.

[7] Marco Fantini, Gabriele Ciravegna, Alkis Koudounas, Tania Cerquitelli, Elena Baralis, Giovanni Succo, and Erika Crosetti. The rapidly evolving scenario of acoustic voice analysis in otolaryngology. *Cureus*, 16, 11 2024.

[8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[9] Yi Lu, Yuankun Xie, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Zhiyong Wang, Xin Qi, Xuefei Liu, Yongwei Li, Yukun Liu, Xiaopeng Wang, and Shuchen Shi. Codecfake: An Initial Dataset for Detecting LLM-based Deepfake Audio. In *Interspeech 2024*, pages 1390–1394, 2024.

[10] Nicolas M. Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2024.

[11] I. Celeste Aurora Solak and Dima Naumov. The m-ailabs speech dataset. https://github.com/i-celeste-aurora/m-ailabs-dataset, 2017.

[12] Vojtěch Staněk, Karel Srna, Anton Firc, and Kamil Malinka. Scdf: A speaker characteristics deepfake speech dataset for bias analysis. In *BIOSIG 2025*. Gesellschaft für Informatik e.V., 2025.

[13] Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. SpeechFake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9985–9998, Vienna, Austria, July 2025. Association for Computational Linguistics.