Acoustic Microcues: Breath and Phoneme-Level Signals for Detecting Audio Deepfakes

Wuqiong Han
lydiahan@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Amogh Palasamudram
amogh.p.214@gmail.com
Georgia Institute of Technology
Atlanta, Georgia, USA

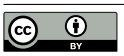
Boo Fullwood
jfullwood3@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Fabian Monrose
fabian@ece.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

Modern voice conversion (VC) and text-to-speech (TTS) systems have reached a point where synthetic speech can no longer be reliably distinguished from genuine voices by human listeners, creating an epistemic challenge for how authenticity is assessed in audio communication. Despite this shift, robust detection methods that operate beyond human perceptual limits or exploit auxiliary signals outside the speech waveform remain scarce. We propose a detection approach grounded in a fundamental constraint of human speech production: in natural speech, the phonetic context surrounding a breath imposes distinct articulatory and aerodynamic demands, giving rise to inhalation timing and spectral patterns that speakers subconsciously coordinate but current deepfake generators do not explicitly model. Building on this observation, we introduce a phoneme-aware, breath-centered detection framework. Using a multi-band spectral centroid extracted from detected breath segments and a lightweight CNN (convolutional neural network) classifier, our method achieves strong performance across state-of-the-art voice conversion systems, with an average F1 score of 0.993 across thousands of test samples. Notably, this performance transfers to a widely deployed commercial generator without retraining, where we achieve an F1 score of 0.949. Beyond accuracy, the simplicity of our model enables transparent feature attribution, providing interpretability absent in large, end-to-end neural detectors and reinforcing that breath–phoneme cues are intrinsically discriminative. The approach further demonstrates robustness under codec transformations commonly used in streaming platforms, with only a 11.37% reduction in F1, and maintains generalization to unseen generators. Finally, to address adversarial attempts that suppress breath, we introduce a tamper detector that flags suspicious samples, providing an additional safeguard for breath-based deepfake detection. It achieves an F1 score of 0.974 in detecting samples from which breath segments have been tampered.

ACM Reference Format:
Wuqiong Han, Amogh Palasamudram, Boo Fullwood, and Fabian Monrose. 2026. Acoustic Microcues: Breath and Phoneme-Level Signals for Detecting Audio Deepfakes. In *2nd Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media (DFF '26), November 10–14, 2026, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3840473.3840514



DFF '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2927-0/2026/11
https://doi.org/10.1145/3840473.3840514

1 Introduction

Deepfake speech refers to synthetically generated audio engineered to mimic human voices, often with striking realism. Indeed, modern generation tools can now produce lifelike speech in minutes, requiring little more than short voice samples and commodity hardware. This progress has unlocked a range of meaningful applications: animators and audiobook creators can craft new voices without lengthy studio sessions; individuals with speech impairments can communicate using voices that reflect their identity; virtual assistants can be tailored to sound more natural or more comforting; and, in therapeutic settings, families can engage with recreated familiar voices as a form of emotional support.

But as this technology became more accessible, its darker potential quickly surfaced. For instance, deepfakes have been weaponized to fabricate political speeches, manipulate public opinion, and disrupt elections. A prominent example occurred just before the 2024 U.S. presidential primaries, when tens of thousands of Democratic voters reportedly received a robocall mimicking President Biden and urging them not to vote. The 40-second fake message was allegedly generated using ElevenLabs' commercial software[1]. In the same year, scammers used deepfake audio and video to steal an estimated $10.7 billion [58].

Despite growing awareness of these threats, human listeners remain poor detectors of synthetic audio. A recent study by [36] involving 529 participants found that listeners failed to identify deepfake speech in over one-third of cases. The situation appears even more pronounced for AI-generated music, where listeners increasingly struggle to distinguish machine-produced tracks from human-created ones.[2] Together, these trends highlight a widening gap between what modern generators can produce and what humans can reliably perceive, underscoring the urgent need for detection methods that operate at a finer level of detail than human judgment alone allows.

[1] V. Balasubramaniyan, "Pindrop Reveals TTS Engine Behind Biden AI Robocall," 2024.
[2] A survey conducted by one of the world's largest independent music streaming platforms (Deezer) reported that 97% of listeners could not reliably distinguish fully AI-generated music from human-created music. See "Clear desire for transparency and fairness for artists," November 12, 2025.

In response, the community has developed a growing arsenal of countermeasures targeting subtle acoustic fingerprints of generative models. Yet most focus on TTS deepfakes, leaving imitation-based voice-conversion (VC) attacks comparatively under-examined. This is concerning: VC models often require less training data, produce more natural and fluent speech [7, 27], and operate with low latency. These advancements make them especially dangerous, and the growing availability of humanlike AI-generated voices has begun to reshape how we evaluate the authenticity of the voices we hear. A second limitation is the limited generalizability and interpretability of existing defenses [2, 60, 61]. Many rely on black-box deep learning models operating on high-dimensional learned representations or abstract low-level acoustic features whose interactions are difficult to interpret. In contrast, we pursue interpretable solutions that, as defined by [14], enable model behavior to be explained in human-understandable terms.

To help bridge these gaps, we propose a pipeline that leverages breath and phonetic cues to detect VC-based deepfakes. Grounded in respiratory and phonetic cues, our approach proves greater transparency and interpretability in decision-making processes and is robust under distributional shift and common signal degradation. The architecture offers clear and intuitive interpretations, making it a practical solution that fills a gap in the current security landscape. This paper makes the following contributions:

(1) We introduce a multi-band spectral centroid feature extracted from breath regions, complemented by phonetic context and grounded in articulatory and respiratory physiology. On realistic datasets with wide speaker coverage, phonetically diverse full-sentence speech, and emotional-prosodic variability, this feature achieves an average F1 of 0.993. Under real-world codec distortions common in VoIP and telephony, it degrades by only 11.37%, substantially smaller than most tested neural detectors, demonstrating that biologically meaningful cues can be as robust as high-dimensional learned embeddings.
(2) We evaluate our method on modern voice-conversion generators, including a widely used commercial system whose outputs are difficult for humans to distinguish from real speech. Even in this out-of-distribution commercial setting, our detector attains an F1 of 0.949, outperforming all competing models and showing that principled feature design generalizes beyond model-specific artifacts.
(3) We develop a tamper detector that identifies whether a sample has been manipulated using breath-suppression attacks, allowing the system to flag adversarial attempts before applying breath-based detection. The tamper detector achieves an F1 score of 0.974 in distinguishing breath-suppressed audio generated using different editing techniques.
(4) We explicitly link detection decisions to phoneme-conditioned respiratory and spectral cues, enabling direct attribution to human-interpretable properties of speech production.

2 Background

Speech is produced as air from the lungs passes through the vocal folds and vocal tract, where it is shaped into speech sounds. Loosely speaking, vowels are produced with an open vocal tract, while consonants involve partial or complete constriction of airflow. Vowels are further characterized by tongue position, lip rounding, and muscular tension, while consonants differ by place and manner of articulation [39].

These articulatory configurations give rise to phonemes, the smallest contrastive sound units in a language; changing one can alter a word's meaning (e.g., the vowel in *lot* versus *lit*). Because phonemes are independent of spelling, they are commonly represented using the International Phonetic Alphabet (IPA), which assigns symbols to distinct speech sounds (e.g., *lot* /lɑt/). While the full IPA represents hundreds of sounds, English varieties such as General American and Received Pronunciation use roughly two dozen consonant and twenty vowel phonemes.

Phonemes differ in their physical articulation and resulting airflow. For example, /p/ (as in puppy) involves complete lip closure followed by a rapid release, producing a sharp burst of airflow. In contrast, /s/ (as in seal) channels air through a narrow constriction, producing sustained turbulence with little outward airflow. Vowels likewise vary in aerodynamic load with tongue position, lip rounding, and vocal tract openness.

Breath events are shaped by adjacent phonemes as the vocal tract transitions between articulation and inhalation. Different phonemes impose distinct articulatory and aerodynamic demands, producing characteristic breath–phoneme interactions. These local interactions capture how individual inhalation events relate to surrounding speech, beyond coarse measures such as breath duration or breath–speech rate. Highly constricted phoneme families (e.g., obstruents, stops, nasals) can particularly influence adjacent breaths. We exploit these interactions to augment breath representations with phonetic context.

3 Related Work

We conducted a systematic review of speech deepfake detection and breath-related signal analysis over the past 15 years, covering Google Scholar, IEEE Xplore, the ACM Digital Library, and arXiv. Search terms and taxonomies were seeded from established audio deepfake surveys and iteratively refined around feature representations, underlying acoustic or physiological signals, and security applications involving breath cues.

Generators. Prior work on speech deepfake generation largely converges on two dominant paradigms: text-to-speech (TTS) and voice conversion (VC) [30]. In the TTS literature, systems are typically described as modular pipelines consisting of a text analysis component that maps orthographic input to linguistic features, an acoustic model that predicts intermediate acoustic representations, and a vocoder that synthesizes the final waveform. By contrast, studies on VC focus on models that operate directly on audio inputs without textual supervision. These systems transform the acoustic features of a source utterance to match those of a target speaker, with the source providing the linguistic content and the target defining speaker-specific characteristics such as timbre and accent. Across the VC literature, implementations vary from explicitly incorporated vocoders to fully end-to-end models that learn the conversion mapping directly.

Recent work in the VC literature mirrors the rapid progress observed in text-to-speech systems. Representative research prototypes include CycleGAN-VC, StarGAN-VC, and more recent models such as FreeVC [29], kNN-VC [3], and OpenVoice2 [45]. These studies consistently emphasize improvements in speech naturalness, speaker similarity, and low-latency or real-time performance. Parallel to academic developments, several commercial platforms, including Speechify, Resemble AI, and ElevenLabs, have made high-quality voice conversion and cloning capabilities broadly accessible. More recently, commercial text-to-speech systems such as ElevenLabs V3 have introduced fine-grained generation controls, enabling explicit manipulation of speaker characteristics and paralinguistic cues, including laughter and breathing.

Detectors. Early advances in text-to-speech synthesis, including WaveNet, Tacotron 2 and YourTTS [10], established benchmarks for speech naturalness and intelligibility and directly motivated the first wave of countermeasure (CM) research. Much of this early work focused on detecting artifacts specific to TTS pipelines, and several approaches achieved strong performance. Notably, RawNet2 [24, 52], a deep neural detector operating directly on raw waveforms, reported state-of-the-art results on cropped text-to-speech samples.

As voice conversion and voice cloning systems became more prevalent, subsequent studies expanded the scope of detection to address these more diverse forms of synthetic speech. Representative approaches include AASIST [23], which applies graph attention mechanisms to feature maps extracted from audio waveforms, and VoiceRadar [26], which leverages Doppler effects and microphone vibration patterns to derive interpretable features with strong performance on VC detection. Other work has explored the identification of VC-specific artifacts by decomposing spectrograms and training detectors on residual or artifact-focused representations [21]. Collectively, these works reflects a shift from TTS-centric defenses toward broader detection strategies designed to capture the heterogeneous characteristics of modern voice conversion systems.

In the past few years, a subset of countermeasure pipelines has begun to incorporate breath-related cues. Drawing on prior work in biometric authentication [11, 12, 62], these approaches build on the observation that certain acoustic properties of breathing exhibit inter-speaker variability while remaining relatively stable over time [11]. Within the deepfake detection literature, [13] proposed BTS-E, a framework that explicitly segments breath and speech events and encodes them as separate representations; when paired with a RawNet2-based classifier, this approach achieved strong performance on TTS samples. Similarly, [28] introduced a breath detection pipeline that leverages temporal breath features for classification, reporting favorable results on a custom dataset constructed from multiple TTS models. However, both lines of work primarily target text-to-speech generation and, by design, do not address more advanced voice conversion or voice cloning systems, which motivates the need for breath-based analyses that extend beyond TTS-centric settings.

Our Work. By leveraging biologically grounded cues, our approach exploits constraints that originate from the mechanics of speech production rather than from generator-specific artifacts. Unlike detectors that operate on entire utterances or rely solely on high-dimensional learned embeddings, our method explicitly aligns inhalation events with phonetic boundaries, enabling contextual analysis of breath–phoneme relationships. This pairing reveals subtle inconsistencies in speech–breath coordination and physiological plausibility that are difficult to capture using conventional speech representations. Moreover, because our features are low-dimensional and interpretable, we can identify which spectral bands and phoneme–breath interactions drive a particular decision. This level of transparency stands in contrast to large end-to-end models that obscure their decision process behind opaque internal representations.

4 Methodology

To train the models, we first performed a systematic assessment of speech data to find a suitable dataset for breath analysis. Using the selected dataset of real speech, we use a set of realistic voice cloning models to produce deepfakes using the sentences and speakers' voices from the bona fide set. The detector models are trained on acoustic features extracted from both the bona fide and deepfake data. At classification time, extracted speech events are used in a classifier for prediction. The overall system pipeline is illustrated in Figure 1. In addition to the primary breath-based deepfake detector, we also introduce a tamper detector module that screens for breath-suppression attacks. We discuss each step in the following subsections.

4.1 Dataset Selection

Although numerous speech datasets exist for ASR and deepfake research, few are suitable for breath-focused analysis. Breathing occurs between longer speech segments and varies with emotional state, yet many widely used datasets consist of short, neutral utterances or crowd-sourced recordings with limited emotional control, making them poor candidates for studying breath behavior.

To identify a suitable dataset, we systematically reviewed English-language speech corpora from academia, industry, government, and leading speech research venues. Selection criteria included: (1) high-quality, low-noise recordings; (2) balanced per-speaker coverage across genders; (3) phonetically diverse, full-sentence utterances long enough to capture multiple breaths; and (4) broad, well-annotated emotional coverage, including both the "Big Six" emotions and more nuanced affective states.

Dataset	Rate (kHz)	#Spk	Gender balance	Emo coverage	Avg length
Librispeech 360 [42]	16	44	✗	✗	12.58
TIMIT [16]	16	630	✗	✗	3.08
Switchboard [17]	8	543	✓	✗	382.66
ESD (Eng.) [63]	16	10	✓	5	2.76
CREMA-D [9]	48	91	✓	6	2.54
ENTERFACE'05 [37]	48	42	✗	6	3.17
EARS [49]	48	107	✓	22	19.81

Table 1: The table lists the sample rate, number of speakers, whether the dataset is gender-balanced, emotion coverage, and the average sample length in seconds.

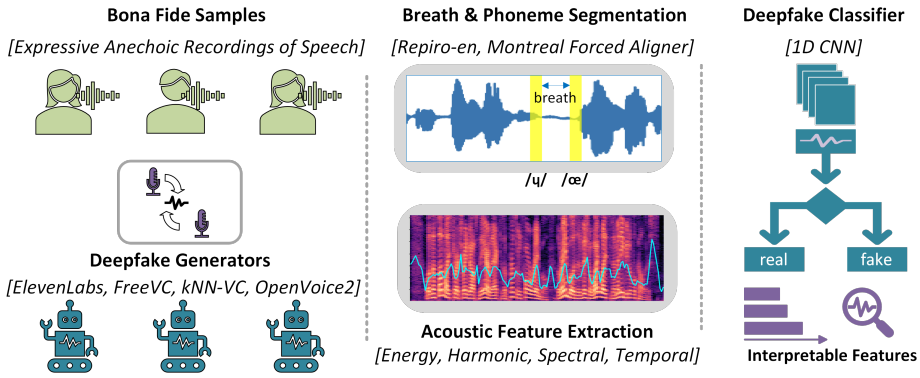


Figure 1: We detect synthetic speech through breath–phoneme coordination. Our approach (1) detects breath segments, (2) aligns them with phonemes, (3) extracts low-dimensional, biologically motivated spectral features, and (4) classifies samples using an interpretable model. These features capture respiratory–articulatory inconsistencies in synthetic speech.

Based on these criteria, the EARS (Expressive Anechoic Recordings of Speech) corpus best satisfies our requirements among the candidate datasets in Table 1. It comprises 107 speakers with diverse demographic and accent backgrounds, provides the broadest emotional coverage of the reviewed datasets, and offers high-fidelity recordings free of background noise. EARS includes emotional speech, read speech, conversations, and non-speech sounds; we use the emotional and read subsets, consisting of 22 emotion-labeled sentences and 32 paragraph excerpts per speaker, respectively.

The diversity of speakers, emotional expressions, and linguistic content, together with suitably long utterance durations, makes the EARS dataset well-suited for breath analysis. Moreover, its clean, high-fidelity recordings help establish a controlled baseline that isolates intrinsic speech properties and avoids confounding factors, before we deliberately introduce distortions such as codecs.

4.2 Generating Deepfakes

From the EARS dataset, we construct a balanced subset of utterance pairs so that every real sample is used at least once as both a source and a target. Since exhaustively pairing of all speakers and utterances would produce millions of conversions across multiple generators, we avoid full enumeration. Our procedure is as follows. Each real utterance is first assigned as a target, with a source selected from a different speaker and content. To maintain balance, we track source usage and prioritize utterances used least frequently, breaking ties randomly. The generators are then applied to pairs of real samples to produce fake samples. We repeat the process with roles reversed so each utterance also appears at least once as a source.

For our main analysis, we use three open-source voice cloning models spanning major VC architectures. FreeVC [29] (2022) is fully neural, while kNN-VC [3] (2023) uses non-parametric nearest-neighbor conversion with neural feature extraction and vocoding. OpenVoice2 [45] (2024) is a TTS-hybrid model that takes audio input while using a TTS backbone to control emotion, accent, and intonation. Together, they span neural, non-parametric, and hybrid paradigms. As a secondary analysis, we evaluate generalizability to ElevenLabs V3, a commercial TTS system offering an ideal attack setting through expressive speech generation with emotional states and explicit breath markers.

4.3 Speech Event Extraction

Our focus is not to develop new phoneme or breath extraction methods, but to analyze how these events aid deepfake classification. Accordingly, we use state-of-the-art tools to detect and label events in each audio sample. For breath detection, we use Respiro-en, a frame-wise model designed to identify breath events in speech [59]. We use the Montreal Forced Aligner (MFA) to align audio with word and phoneme boundaries, providing state-of-the-art accuracy with low computational overhead [35, 38]. Other aligners, such as WhisperX, could be substituted without affecting our approach. We validate detected breath regions against established human respiration benchmarks. Prior work reports an average breath-group duration of 0.54 s [55] and maximum breath zero-crossing rate below 0.3 [50]. Both real and synthetic speech in our data fall within these ranges, indicating plausible acoustic and temporal breath characteristics.

4.4 Feature Design

We use a multi-band spectral centroid, a vector feature that measures the weighted average frequency of the power spectrum in multiple frequency bands. The spectral centroid has long been used in acoustics to quantify the perceived brightness of musical instruments [51] and to separate or recognize vowel sounds [22]. But instead of computing a single centroid over the full frequency range, we divide the spectrum into multiple bands and compute a centroid within each band, visualized in Figure 2. The benefit of the multi-band approach is that it allows us to capture localized spectral details that the full-spectrum centroid would obscure. To the best of our knowledge, we are the first to apply the multi-band centroid to a deepfake detection setting.

We compute the spectral centroid for each breath window as

$$\text{Centroid}(t) = \frac{\sum_k f_k |S(f_k, t)|}{\sum_k |S(f_k, t)|}, \tag{1}$$

where $S(f_k, t)$ denotes the short-time Fourier transform magnitude at frequency bin k and time t.

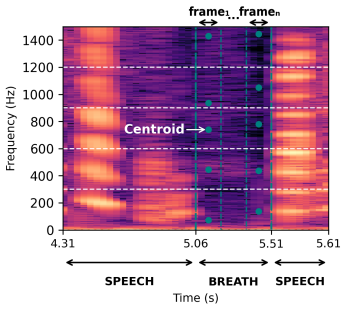


Figure 2: Illustration of a multi-band centroid. Each centroid represents the weighted mean frequency within its corresponding band. For each frame, the set of band-wise centroids forms a feature vector.

To demonstrate the value of this feature in the domain of deepfake detection, we compare it against a range of established signal-processing measures used in speech and general acoustic analysis. These baseline features were selected through a systematic review of classic signal processing textbooks [18, 46], speech-focused works emphasizing hand-crafted or rule-based features [5, 20, 44, 59], and surveys on features used in deepfake detection [2, 60, 61]. The features span temporal, energy, spectral, and harmonic categories, and have been used in applications such as tamper detection, authentication, and noise analysis[48, 61].

Spectral bandwidth measures how energy is distributed around the spectral centroid. Root mean square (RMS) energy reflects perceptual loudness and is commonly used to characterize both speech and breath signals. Fundamental frequency (F_0), the lowest frequency of a waveform, corresponds closely to perceived pitch and is a strong descriptor of harmonic structure.

Intuitively, the Harmonics-to-Noise Ratio (HNR) quantifies the balance between periodic and aperiodic energy and is widely used to assess voice quality. Similarly, the zero-crossing rate (ZCR), which measures the number of times the signal crosses zero amplitude, has proven effective for breath detection [59] and for distinguishing voiced from unvoiced sounds [4]. We also include utterance-level duration-based features that capture coarse temporal breathing patterns (e.g., breath duration and breath–speech rate) that have been leveraged in prior deepfake detection work [8, 28].

Finally, a small number of biometric security studies employ sub-band analysis over Mel-Frequency Cepstral Coefficients (MFCCs). Although this band-bucketing strategy is conceptually related to our approach, MFCCs are highly derived representations that summarize the spectral envelope and discard fine-grained detail, limiting their interpretability in physiological or articulatory terms. We therefore exclude this class of features from the evaluation in Section 5. Our focus is instead on low-level, human-interpretable features that preserve a direct connection to respiratory acoustics and support transparent attribution of detection decisions.

Feature extraction follows established conventions [31]. Audio is resampled to 16kHz with a 1024-point FFT and 256-sample hop, yielding a 16 ms frame step, well below the ∼200 ms lower bound of average breath duration [55]. For non-duration features, we compute summary statistics across frames within each breath window as classifier inputs.

4.5 Metrics

We evaluate detection performance using two standard metrics: F1 score and equal error rate (EER). F1 balances precision and recall, with deepfakes as the positive class and bona fide samples as negative. We use a fixed decision threshold of 0.5. EER is the threshold-independent point where the false acceptance rate (FAR; deepfakes classified as bona fide) equals the false rejection rate (FRR; bona fide samples classified as fake); lower EER indicates better discriminability.

5 Evaluation

From the broader landscape of CMs discussed in Section 3, we select four representatives for comparison: Wav2Vec2.0 [6], AST [19], Whisper [47], and Layton et al.[28]. Wav2Vec2.0 underpins many modern deepfake detection and ASR pipelines [53, 56, 57]. AST is a spectrogram-based transformer that has demonstrated strong accuracy and robustness under distribution shift [15, 54]. Whisper is a widely adopted, large-scale model shown to perform well across tasks, including deepfake detection and speech emotion recognition [1, 30, 41]. Collectively, these three models represent deepfake detection approaches adapted from large pretrained models, relying on complex architectures, high-dimensional learned representations, and large-scale training data. In contrast, Layton et al. use coarse-grained temporal breath statistics, providing a direct breath-based comparison. We evaluate our biologically informed, lightweight approach against these baselines to assess its effectiveness relative to both complex learned detectors and an existing breath-based approach. To ensure fairness, we finetune all competitor detectors to our data.

To demonstrate the effectiveness of our multi-band centroid feature, we evaluate it using both a lightweight CNN and a tree-based XGBoost (eXtreme Gradient Boosting) classifier. We intentionally choose these setups to keep the models simple while allowing the proposed feature representation to remain the primary source of discriminative information. Unlike large end-to-end neural detectors that rely on high-dimensional learned embeddings, our shallow models operate on low-dimensional, physiologically-grounded features, enabling easier attribution of decisions to specific acoustic bands or breath–phoneme interactions.

Our CNN uses two 1D convolutional blocks (kernel size 2) followed by dropout (0.3), and is trained with Adam and cross-entropy loss for up to 50 epochs with early stopping. XGB uses 300 boosting rounds, a maximum depth of 6, a learning rate of 0.05, and logistic output. Across experiments, we use real-fake balanced 60:20:20 train–validation–test splits with no sample overlap.

We first examine the discriminative contribution of individual feature families. Table 2 shows the classification results using the advanced voice-cloning models discussed in Section 4.2 on the EARS dataset. Spectral features are the top-performers, with the multi-band centroid consistently providing the strongest discriminative power across all VC models. Coarse temporal features, by contrast,

Feature	OV2	FVC	KNN	Avg F1
Multi-band Centroid	0.9871	0.9916	0.9996	**0.9930**
RMS Energy	0.8708	0.9824	0.9337	0.9290
Spectral Bandwidth	0.9214	0.8827	0.9546	0.9196
Full Spectrum Centroid	0.9063	0.8098	0.8938	0.8699
ZCR	0.8900	0.6350	0.7891	0.7714
F_0	0.6624	0.6873	0.6474	0.6657
Breath Rate	0.5489	0.5183	0.5676	0.5449
HNR	0.4375	0.5797	0.3935	0.4702
Breath Duration	0.4854	0.1478	0.4903	0.3745

Table 2: F1 performance when trained on each feature individually, across VC systems (OV2 = OpenVoice2, FVC = FreeVC, KNN = kNN-VC). Best and second-best performing features are highlighted in green and orange, resp. Features are sorted by average F1.

perform no better than random guessing. Given the near-saturated performance of the multi-band centroid here, subsequent evaluations include more challenging settings to examine the robustness and forensic interpretability of added phoneme features alongside the centroid.

5.1 Deepfake Classification

Given the strong performance of the multi-band centroid feature extracted from detected breath regions, we compare it against three leading neural deepfake detectors (Wav2Vec2.0, AST, Whisper) [19, 31, 32], as well as the breath-based approach of Layton et al. [28].

Our first evaluation considers each generator (FreeVC, kNN-VC, or OpenVoice2) independently to assess per-generator performance (Table 3). Across all VC generators, our method achieves an average F1 of 0.993 (EER = 0.007). Even for the lowest-performing case (OpenVoice2), F1 reaches 0.988. The tree-based XGB classifier similarly achieves strong performance. By comparison, Layton et al.'s coarse breath statistics performs poorly on VC-based deepfakes, highlighting the non-trivial role of feature design and granularity in representing breath signals. Overall, our lightweight models perform on par with three large neural approaches relying on large-scale training data and highly parameterized features, while substantially outperforming this closest breath-based baseline. The simplicity and stability of our method complement the strong discriminative power of our features, demonstrating that, contrary to common wisdom, accurate deepfake detection does *not* require complex neural architectures when physiologically grounded cues are properly extracted.

5.1.1 Codec Impacts. While our performance is slightly lower than SOTA neural approaches, the strength of our multi-band centroid feature is more evident under real-world audio conditions, particularly when deepfake signals are processed through coder-decoders (codecs) used in streaming platforms and telephony. Such algorithms compress audio by discarding or approximating components deemed less perceptible to human hearing, often introducing distortions that can severely degrade detector performance. To assess robustness under these practical constraints, we evaluate all detectors on deepfakes degraded by the Opus Narrowband codec (8 kHz sample rate and 8kbit/s bitrate), commonly used in VoIP and streaming applications. The use of the codec allows us to simulate lossy transmission conditions under which many real-world deepfake attacks occur, allowing us to test whether breath-centric cues remain stable even after aggressive compression.

Detector	OpenVoice2	FreeVC	KNN-VC	Avg F1
Ours (CNN)	0.9880	0.9912	0.9994	0.993
Ours (XGB)	0.9810	0.9870	0.9920	0.987
Wav2Vec	1.0000	1.0000	1.0000	1.000
AST	1.0000	1.0000	1.0000	1.000
Whisper	1.0000	0.9950	1.0000	0.998
Layton et al.	0.5500	0.5500	0.5400	0.547

Table 3: Detector performance across VC generators.

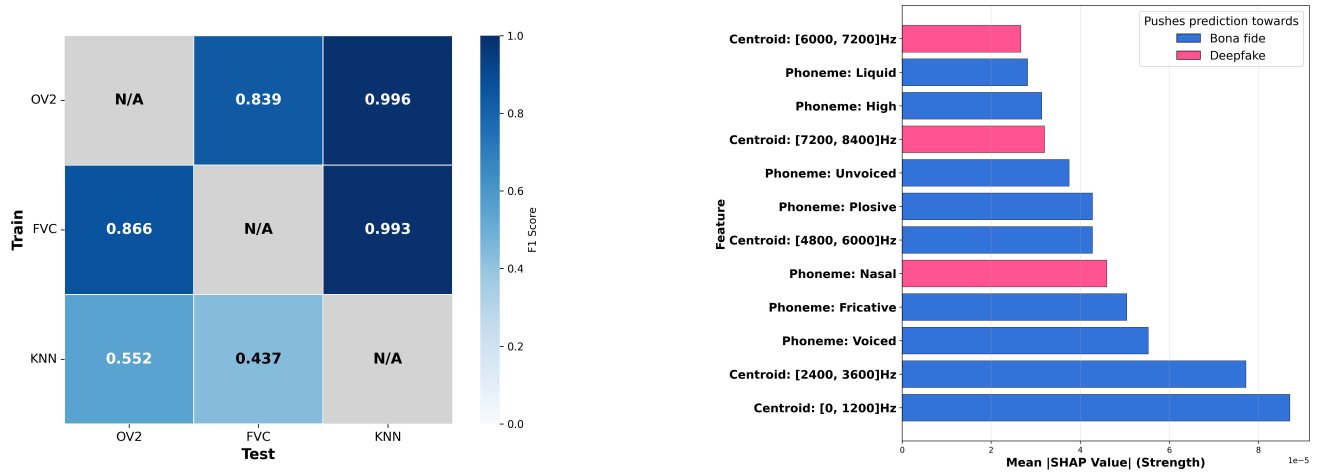
The performance comparison under this setup is shown in Table 4. In addition to raw F1 scores, we report performance changes relative to clean audio to capture model robustness. Under codec degradation, both of our lightweight approaches remain robust, with the CNN showing a modest 11.37% decline in F1 (EER +0.02), substantially smaller than the degradation observed for two state-of-the-art neural detectors. Wav2Vec and AST, both of which achieve perfect accuracy under clean conditions, suffer dramatic declines of 36.47% and 22.97%, respectively. Although Whisper exhibits the strongest resilience overall, this outcome is unsurprising given that its architecture is trained with aggressive spectro-temporal augmentation [43], which mimics distortions introduced by codec transformations.

Approach	Baseline F1 (Avg.)	Codec F1 (Avg.)	ΔF1 (%)
Whisper	0.9983	0.9223	-7.60
Ours (CNN)	0.9926	0.8790	-11.37
Ours (XGB)	0.9984	0.8597	-13.87
AST	1.0000	0.7703	-22.97
Wav2Vec	1.0000	0.6353	-36.47
Layton et al.	Excluded due to low baseline performance		

Table 4: F1 under Opus narrowband codec degradation. Best and second-best robustness (smallest ΔF1) are highlighted.

Notably, our lightweight model, built on physiologically grounded breath–phoneme features rather than high-capacity neural embeddings, performs remarkably well in comparison. Despite its simplicity, it remains robust to real-world degradations that cause substantially larger performance drops in more complex neural systems. This highlights the practical value of biologically informed feature design, particularly under compression and bandwidth constraints.

5.1.2 Generalization. To approximate real-world deployment, we also evaluate how well the detectors generalize to unseen deepfake generators. Here, we train on fake audio produced by one generator and test on fake samples produced by a different generator, while all bona fide audio is drawn from the EARS dataset. Train, validation, and test splits remain strictly disjoint in terms of sample, yielding the six cross-generator combinations shown in Figure 3a.



(a) Cross-generation performance heatmap.

(b) SHAP feature importance analysis.

Figure 3: (Left) Cross-generation F1 performance heatmap (OV2 = OpenVoice2, FVC = FreeVC, KNN = kNN-VC). (Right) SHAP analysis of the proposed model. A feature's magnitude reflects how strongly it influences the classifier's prediction, while its color indicates whether it pushes the prediction toward the deepfake (positive) or bona fide (negative) class.

Our CNN model demonstrates good robustness in most settings, particularly when trained on OpenVoice2. The robustness (0.92 F1, 0.04 EER average) likely stems from the fact that OpenVoice was specifically designed to decouple the components in a voice (e.g., the generation of language, tone, and other important style parameters like emotion and rhythm) as much as possible. This is in sharp contrast to the case where we train on kNN-VC. Unlike the neural VC models, kNN-VC converts each frame by averaging its k nearest neighbors in the target speaker's feature space, producing frames with unusually low variance. These highly uniform patterns make kNN-VC trivially detectable within its own domain but offer little transferability to the richer dynamics exhibited by other generators. Notably, the phoneme features provide a key benefit in the cross-generator setting by improving overall F1. The improvement is most pronounced in the challenging kNN-based setting, where phoneme features raise the average cross-generator F1 from 0.44 (centroid-only) to 0.49. This suggests that phonetic context strengthens the performance floor under distribution shift, with its benefits becoming more apparent beyond the near-saturated in-distribution setting.

5.2 Case Study

We further evaluate our approach on deepfakes generated by ElevenLabs, a widely used commercial system whose outputs are difficult for human listeners to distinguish from real speech [7, 27]. Recent ElevenLabs models also support fine-grained generation control, including deliberate breath insertion and adaptation of prosodic characteristics such as tone, intonation, and speaking rate, presenting a challenging real-world setting.

We evaluate 370 samples generated with Eleven V3, its latest text-to-speech model, using EARS transcripts. For each sample, we randomly select a gender-matched recommended voice and include an emotion tag (e.g., cutely" or sarcastically") when available. Breath commands are inserted at locations detected in the original speech and specified as short, medium, or long based on the corresponding duration. All other generation settings remain at their defaults.

Despite this challenging real-world setting, our physiologically grounded feature set remains highly effective, achieving an F1 score of 0.949 (EER = 0.04). Our method achieves the highest F1 among the evaluated detectors (next-best F1: AST = 0.91; EER = 0.03). This highlights a key result of our work: *carefully chosen, biologically motivated features can reveal subtle inconsistencies in synthetic audio, even when the generator produces speech that sounds natural to human listeners.*

5.3 Suppression Attacks and Tamper Detection

An obvious attack in this context is breath suppression. We evaluate two scenarios under a white-box adversary who is aware of our breath-based detector and attempts to suppress breath evidence. In the first, the attacker approximates breath regions and removes all but one randomly selected breath, leaving minimal breath information. Even under this aggressive attack, our detector achieves an average F1 of 0.960, only a 3.3% drop from 0.993 with all breaths preserved, demonstrating strong discriminative performance from a single remaining breath event.

In the second scenario, the attacker completely suppresses all breath events, eliminating the primary evidence for breath-based detection. We address this threat with a tamper detector that identifies breath-suppressed inputs before classification. During deployment, inputs deemed untampered proceed to the primary breath-based detector, while suspicious inputs are deferred to complementary methods, such as speech-based detectors, whose design is outside the scope of this work.

Our tamper detector uses temporal centroid differences between neighboring frames rather than absolute values, as breath suppression can introduce abrupt spectral discontinuities at edit boundaries. We construct the tampered class using three common operations: zeroing breath regions, removing and directly concatenating breath segments, and removal followed by crossfaded concatenation [25, 34]. These attacks assume moderate signal-processing knowledge for frame-level audio manipulation. The detector is trained on balanced tampered and untampered samples containing both bona fide and deepfake audio, achieving an average F1 of 0.974 in identifying manipulated inputs before breath-based analysis.

6 Interpretability

To instill confidence that the breath-region features drive classification performance, we use SHapley Additive exPlanations (SHAP) [33], a widely adopted framework for interpreting model predictions. SHAP quantifies each feature's contribution by computing its marginal effect averaged over all possible feature subsets. Figure 3b shows the mean SHAP value of different features across test predictions for our detector, which achieves an average F1 of 0.993. The horizontal axis shows the magnitude of the SHAP effect, while color encodes its direction: red indicates contributions that push the prediction toward the fake class, and blue indicates contributions that push the prediction toward the bona fide class.

The centroid bands in Fig. 3b shows that the 0–1200Hz and 2400–3600Hz regions exhibit the largest SHAP magnitudes toward the bona fide class, indicating that these low-frequency cues contribute most strongly to the model's decisions. At the higher end of the spectrum, the 6000–7200Hz and 7200–8400Hz bands exhibit a notable contribution toward the fake class. This band-dependent importance aligns with the known spectral characteristics of human breathing. Prior work [40] has shown that breath sounds concentrate most of their energy at lower frequencies, particularly below 150Hz. Consequently, low-frequency centroid features provide the most informative respiratory cues, while higher-frequency bands, although less consistently present, may capture complementary artifacts introduced by deepfake generation.

While centroid features provide the primary discriminative signal, the phoneme features in Fig. 3b provide complementary linguistic context by characterizing articulation surrounding each breath. Among them, adjacent voiced phonemes (including vowels) show the strongest predictive power. Their sustained voicing and high sonority create acoustically distinct breath–voice transitions, providing strong discriminative cues [39]. Nasals (e.g., /m/, /n/) contribute strongly toward the fake class, suggesting that generators may poorly reproduce breath–nasal transitions. Plosives (e.g., /p/) contribute strongly toward real predictions, potentially reflecting their precise articulatory coordination and characteristic speech–breath transitions. Liquids (e.g., /l/), with complex tongue configurations and strong co-articulation, may likewise pose challenges for deepfake generators.

7 Limitations

Although breath–phoneme cues provide a strong discriminator for current deepfake systems, several factors limit their universal applicability. First, breath sounds can be attenuated or masked by environmental noise and aggressive audio compression. Although our feature design remains robust under common streaming and VoIP codecs (Section 5.1.1), more extreme compression or poor recording conditions may further degrade breath detectability [59]. Alternatively, a motivated adversary could insert prerecorded breaths or manually adjust breath timing to better mimic human respiratory patterns. While our tamper detector raises the bar for these attacks by identifying artifacts introduced by breath-suppression manipulations, detecting adversarial breath insertion or stitching would likely require complementary analyses of signal continuity and acoustic naturalness. This limitation, however, is shared by all stand-alone breath-based security solutions [11, 13, 28, 62]. Dataset diversity is another limitation. EARS includes demographic diversity and emotional and prosodic variation crucial for capturing breath–speech dynamics. Comparable datasets remain scarce. Future work will investigate additional datasets and speaker-independent evaluation.

Overall, our use of breath-based cues offers a valuable complementary signal within a defense-in-depth strategy. By targeting constraints of human speech production that current generative systems do not explicitly model, they provide interpretable evidence that can aid forensic analysis and model auditing. Accordingly, we argue that breath–phoneme features should be viewed not as a standalone solution, but as a low-complexity, transparent component that strengthens broader synthetic speech detection pipelines.

8 Conclusion

Audio deepfake detection has followed a similar trajectory to other ML-focused tasks wherein advanced speech generators necessitate more advanced and complex detectors. This has been borne out in both improved detection performance and in ballooning model parameters, feature extraction complexity, and pretraining requirements. However, our success in achieving near state-of-the-art detection performance across a range of modern speech generators demonstrates that more model complexity is not necessary and can often be replaced with careful selection of domain-specific features. Simple spectral analysis, augmented with an understanding of the physiological characteristics of speech production and its impact on breath, allows us to maintain competent detection (0.993 F1), even on out-of-distribution data and the most advanced generators. Additionally, these features tolerate signal degradation through lossy encoding better than the majority of learned features (11.37% performance loss vs. 29.72% loss) unless the model is specifically hardened against this distortion.

9 Ethics, Open Science, and AI Disclosure

The evaluated VC generators are publicly available research implementations or commercial systems used in accordance with their terms of service. For reproducibility, we release our implementation at[3]. We believe the benefits of transparency outweigh misuse risks, as our release introduces no new generative capabilities. AI use was limited to cosmetic edits.

[3]https://doi.org/10.5281/zenodo.22007137

References

[1] Aulia Adila, Dessi Lestari, Ayu Purwarianti, Dipta Tanaya, Kurniawati Azizah, and Sakriani Sakti. 2024. Enhancing Indonesian Automatic Speech Recognition: Evaluating Multilingual Models with Diverse Speech Variabilities. In *27th Conference of the Oriental International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques*. 1–6.

[2] Zaynab Almutairi and Hebah Elgibreen. 2022. A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions. *Algorithms* 15, 5 (2022).

[3] Matthew Baas, Benjamin van Niekerk, and Herman Kamper. 2023. Voice conversion with just nearest neighbors. *arXiv preprint arXiv:2305.18975* (2023).

[4] R Bachu, BK Adapa, S Kopparthi, and BD Barkana. 2008. Separation of Voiced and Unvoiced Speech Signals using Energy and Zero Crossing Rate. In *ASEE Regional Conference, West Point*.

[5] R.G. Bachu, S. Kopparthi, B. Adapa, and B.D. Barkana. 2010. Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. In *Advanced Techniques in Computing Sciences and Software Engineering*. 279–282.

[6] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*. 12449–12460.

[7] Sarah Barrington, Emily A. Cooper, and Hany Farid. 2025. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports* 15, 1 (2025), 11004.

[8] Gila Benchetrit. 2000. Breathing pattern in humans: diversity and individuality. *Respiration Physiology* 122, 2 (2000), 123–129.

[9] Houwei Cao, David Cooper, Michael Keutmann, Ruben Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5 (10 2014), 377–390.

[10] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In *Proceedings of the 39th International Conference on Machine Learning*. 2709–2720.

[11] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing Acoustics-based User Authentication. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 278–291.

[12] Jagmohan Chauhan, Suranga Seneviratne, Yining Hu, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2018. Breathing-Based Authentication on Resource-Constrained IoT Devices using Recurrent Neural Networks. *Computer* 51, 5 (2018), 60–67.

[13] Thien-Phuc Doan, Long Nguyen-Vu, Souhwan Jung, and Kihun Hong. 2023. BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[14] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[15] Boo Fullwood and Fabian Monrose. 2026. Seeing Is Believing: Interpreting Behavioral Changes in Audio Deepfake Detectors Arising from Data Augmentation. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security*. 206–217.

[16] J. Garofolo, Lori Lamel, W. Fisher, Jonathan Fiscus, D. Pallett, N. Dahlgren, and V. Zue. 1992. TIMIT Acoustic-phonetic Continuous Speech Corpus. *Linguistic Data Consortium* (11 1992).

[17] J Godfrey and E Holliman. 1997. Switchboard-1 release 2: Linguistic data consortium. *Switchboard: a user's manual* (1997).

[18] Ben Gold, Nelson Morgan, and Dan Ellis. 2011. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons.

[19] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[20] Fabien Gouyon, Francois Pachet, and Olivier Delerue. 2002. On the Use of Zero-Crossing Rate for an Application of Classification of Percussive Sounds. (08 2002).

[21] R. Hemavathi and R. Kumaraswamy. 2021. Voice conversion spoofing detection by exploring artifacts estimates. *Multimedia Tools Appl.* 80, 15 (June 2021), 23561–23580.

[22] Risanuri Hidayat, Deni Yulian, Agus Bejo, and Sujoko Sumaryono. 2017. Comparison of Vowel Feature Extraction on Time and Frequency Domain. In *2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE)*. 1–4.

[23] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6367–6371.

[24] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu. 2020. Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *arXiv preprint arXiv:2004.00526* (2020).

[25] Nicholas Klein, Hemlata Tak, James Fullwood, Krishna Regmi, Leonidas Spinoulas, Ganesh Sivaraman, Tianxiang Chen, and Elie Khoury. 2025. Pindrop it! Audio and Visual Deepfake Countermeasures for Robust Detection and Fine Grained-Localization. arXiv:2508.08141 [cs.CV] https://arxiv.org/abs/2508.08141

[26] Kavita Kumari, Maryam Abbasihafshejani, Alessandro Pegoraro, Phillip Rieger, Kamyar Arshi, Murtuza Jadliwala, and Ahmad Sadeghi. 2025. VoiceRadar: Voice Deepfake Detection using Micro-Frequency and Compositional Analysis. *Proceedings 2025 Network and Distributed System Security Symposium* (2025).

[27] Nadine Lavan, Mairi Irvine, Victor Rosi, and Carolyn McGettigan. 2025. Voice clones sound realistic but not (yet) hyperrealistic. *PLOS ONE* 20, 9 (2025), 1–19.

[28] Seth Layton, Thiago De Andrade, Daniel Olszewski, Kevin Warren, Carrie Gates, Kevin Butler, and Patrick Traynor. 2025. Every Breath You Don't Take: Deepfake Speech Detection Using Breath. *Digital Threats* 6, 3, Article 13 (Sept. 2025).

[29] Jingyi Li, Weiping Tu, and Li Xiao. 2023. Freevc: Towards High-Quality Text-Free One-Shot Voice Conversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[30] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2025. A Survey on Speech Deepfake Detection. *ACM Comput. Surv.* 57, 7, Article 165 (Feb. 2025).

[31] Xiang Li, Pin-Yu Chen, and Wenqi Wei. 2025. Measuring the robustness of audio deepfake detectors. *arXiv preprint arXiv:2503.17577* (2025).

[32] Xiang Li, Pin-Yu Chen, and Wenqi Wei. 2025. Where are We in Audio Deepfake Detection? A Systematic Analysis over Generative and Detection Models. *ACM Trans. Internet Technol.* 25, 3, Article 20 (Aug. 2025).

[33] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[34] Hieu-Thi Luong, Haoyang Li, Lin Zhang, Kong Aik Lee, and Eng Siong Chng. 2025. LlamaPartialSpoof: An LLM-Driven Fake Speech Dataset Simulating Disinformation Generation. arXiv:2409.14743 [eess.AS] https://arxiv.org/abs/2409.14743

[35] Tristan J. Mahr, Visar Berisha, Kan Kawabata, Julie Liss, and Katherine C. Hustad. 2021. Performance of Forced-Alignment Algorithms on Children's Speech. *Journal of Speech, Language, and Hearing Research* 64, 6S (2021), 2213–2222.

[36] Kimberly T. Mai, Sergi Bray, Toby Davies, and Lewis D. Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE* 18, 8 (Aug. 2023), e0285333.

[37] O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006. The eNTERFACE'05 Audio-Visual Emotion Database. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*. 8.

[38] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi.. In *Interspeech*. 498–502.

[39] W. O'Grady, J. Archibald, M. Aronoff, and J. Rees-Miller. 2004. *Contemporary Linguistics: An Introduction*. Bedford/St. Martin's. https://books.google.ne/books?id=ObC_QgAACAAJ

[40] Ana Oliveira and Alda Marques. 2014. Respiratory sounds in healthy people: A systematic review. *Respiratory Medicine* 108, 4 (2014), 550–570.

[41] Mohamed Osman, Daniel Z Kaplan, and Tamer Nadeem. 2024. Ser evals: In-domain and out-of-domain benchmarking for speech emotion recognition. *arXiv preprint arXiv:2408.07851* (2024).

[42] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210.

[43] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019 (interspeech_2019)*. ISCA, 2613–2617.

[44] Geoffroy Peeters et al. 2004. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Ist Project Report* 54, 0 (2004), 1–25.

[45] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479* (2023).

[46] L.R. Rabiner and R.W. Schafer. 2007. *Introduction to Digital Speech Processing*. Now. https://books.google.com/books?id=Z6Otr8Hj1WsC

[47] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. 28492–28518.

[48] Douglas A. Reynolds. 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17, 1 (1995), 91–108.

[49] Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. 2024. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. *arXiv preprint arXiv:2406.06185* (2024).

[50] Dima Ruinskiy and Yizhar Lavner. 2007. An Effective Algorithm for Automatic Detection and Exact Demarcation of Breath Sounds in Speech and Song Signals. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 3 (2007), 838–850.

[51] Emery Schubert, Joe Wolfe, and Alex Tarnopolsky. 2004. Spectral centroid and timbre in complex, multiple instrumental textures.

[52] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-End anti-spoofing with RawNet2. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

6369–6373.

[53] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233* (2022).

[54] Andrew Ustinov, Matey Yordanov, Andrei Kuchma, and Mikhail Bychkov. 2025. Cross-technology generalization in synthesized speech detection: Evaluating ast models with modern voice generators. *arXiv preprint arXiv:2503.22503* (2025).

[55] Yu-Tsai Wang, Jordan R Green, Ignatius SB Nip, Ray D Kent, and Jane Finley Kent. 2010. Breath group analysis for reading and spontaneous speech in healthy adults. *Folia phoniatrica et logopaedica* 62, 6 (2010), 297–302.

[56] Zhiyong Wang, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Xiaopeng Wang, Yuankun Xie, Xin Qi, Shuchen Shi, Yi Lu, Yukun Liu, Chenxing Li, Xuefei Liu, and Guanjun Li. 2025. Mixture of Experts Fusion for Fake Audio Detection Using Frozen wav2vec 2.0. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 1–5.

[57] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2023. Learning a Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection. In *Interspeech 2023*. 2808–2812.

[58] Yahoo! Finance. 2024. AI-Fueled Crypto Scams Are Booming. https://finance.yahoo.com/news/ai-fueled-crypto-scams-booming-172752391.html.

[59] Dong Yang, Tomoki Koriyama, and Yuki Saito. 2024. Frame-wise breath detection with self-training: An exploration of enhancing breath naturalness in text-to-speech. *arXiv preprint arXiv:2402.00288* (2024).

[60] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970* (2023).

[61] Bowen Zhang, Hui Cui, Van Nguyen, and Monica Whitty. 2025. Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead. *Sensors* 25, 7 (2025).

[62] Wenbo Zhao, Yang Gao, and Rita Singh. 2017. Speaker identification from the sound of the human breath. *arXiv preprint arXiv:1712.00171* (2017).

[63] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication* 137 (2022), 1–18.