RealComm: Benchmarking and Adapting Audio Deepfake Detection over Real Mobile-Call Channels

***Abstract*—Audio deepfake detectors are often evaluated on digital waveforms, whereas a detector operating on received call audio encounters the combined effects of acoustic capture, device processing, and communication transmission. We introduce RealComm, a paired corpus for examining this discrepancy under real mobile-call conditions. A balanced bilingual pool of 2,800 bona fide and synthetic utterances supports source-disjoint partitions, with synthetic speech from seven generation systems. RealComm-Train contains 12,320 recordings for training and development under 44 conditions, and RealCommBench contains 23,520 test recordings under 42 conditions. The acquisition design varies acoustic versus wired injection, handset versus speakerphone operation, directed device routes, and user-controlled denoising. Across six existing detectors, call transmission increases equal error rate (EER) by 11.73–37.95 percentage points. Acoustic injection is consistently more difficult than wired injection. Paired analyses relate this degradation to recognition errors, speaker similarity, predicted speech quality, and spectral changes; a reduction in high-frequency energy alone does not explain the ordering of detection difficulty. We further develop RealComm-Aug, a staged waveform augmentation pipeline, and compare simulated augmentation, real-call adaptation, and their combination on three detectors. Combined adaptation yields the lowest overall call EER for each model, reaching 22.37% for Teffic-Audio, compared with 37.94% before adaptation. Improvements extend to the held-out device and unseen route combinations, while acoustic injection and condition-dependent tradeoffs remain challenging.**

***Index Terms*—Audio deepfake detection, communication channels, speech anti-spoofing, domain shift, data augmentation, real-world benchmark.**

I. Introduction

RECENT speech generation systems produce convincing speech from text and a short reference recording [1]–[3]. This progress makes the communication path an important part of audio deepfake evaluation. In a phone call, a generated waveform may enter the caller's device through a cable or through a loudspeaker and microphone. Device processing and communication transmission then modify the signal before it reaches the receiver. A detector applied to that received waveform consequently operates on a different signal distribution from the one represented by the source file.

Successful detection in the digital domain does not establish reliability after this transformation. A model may rely on fine spectral or temporal details that are altered by capture, filtering, or encoding. At the same time, received speech can remain intelligible and preserve the intended speaker identity. Detection quality and perceived speech quality therefore need to be measured separately. The practical question is not simply whether a channel changes the waveform, but which changes accompany detector failures and which adaptation strategies improve performance on real calls.

Answering this question requires paired observations. Comparing unrelated digital and call datasets confounds the communication path with content, speaker, and synthesis-system differences. Purely simulated corruptions are useful for controlled experiments, but cannot establish that a model generalizes to the combination of processing steps present in a real device-to-device call. We construct RealComm around repeated acquisition of the same source utterances. Both bona fide and synthetic speech traverse the same set of recording conditions, and every call recording retains a link to its digital source.

The acquisition protocol covers three input configurations, two partially overlapping device rings, and the caller's user-controlled denoising setting. The rings provide a manageable recording design in which every device acts as both caller and receiver. Their overlap supports separate analyses of previously covered routes, new combinations of covered devices, and a device reserved for testing. Source-level partitioning keeps all versions of an utterance together and avoids treating repeated channel realizations as independent linguistic content.

We study three questions. First, how much do existing detectors degrade after real call transmission, and how does this depend on the input path and processing conditions? Second, do measured quality and spectral changes track detection degradation? Third, how much can simulated augmentation and limited real-call training recover, individually and together? The contributions are as follows:

- A paired bilingual real-call corpus with 12,320 training/development recordings and 23,520 test recordings, accompanied by source-level partitions and condition metadata.
- A six-model zero-shot evaluation and paired quality analysis that expose a large digital-to-call gap, the difficulty of acoustic injection, and the limits of explaining failures by high-frequency attenuation alone.
- A staged augmentation pipeline and a three-model adaptation study showing complementary gains from simulated perturbations and real-call data, together with their remaining condition-dependent limitations.

II. Related Work

A. Benchmarks and Generalization

Audio deepfake evaluation has expanded from controlled spoofing challenges to more varied attacks and acquisition settings. ASVspoof 2019 distinguishes logical-access synthesis and conversion from physical-access replay [4]. ASVspoof 2021 adds transmission and compression variation [5], while ASVspoof 5 broadens speaker, recording, and attack diversity through crowdsourced resources [6]. Its evaluation further

highlights vulnerability to adversarial attacks and neural compression [7]. These developments make the test distribution central to interpreting detector performance.

Complementary datasets vary the source material and generation mechanisms. WaveFake provides controlled synthetic-speech data across neural waveform generators [8]. MLAAD emphasizes multilingual synthesis [9], and Codecfake addresses speech associated with neural audio codecs [10]. In-the-wild evaluation reveals poor transfer from established benchmarks to collected online speech [11]; SpoofCeleb incorporates diverse real-world source conditions into speech generation and detection evaluation [12]. A broader survey likewise identifies generalization as a continuing challenge [13]. These resources motivate broader coverage, but do not make source variation and transmission variation interchangeable. RealComm holds the source utterance fixed while varying its real acquisition path.

B. Channel Effects and Real Communication

Zhang et al. show that channel mismatch can degrade spoofing countermeasures and examine augmentation, multi-task learning, and adversarial learning as remedies [14]. More general corruption studies consider noise, signal modification, and compression, finding different vulnerabilities across detector families [15]. Their controlled perturbations help identify sensitivities, whereas a real call combines several processing stages whose individual settings may not be exposed.

Communication-focused benchmarks already address this deployment gap. ADD-C evaluates codec compression and simulated packet loss and studies communication-oriented augmentation [16]. RTCFake examines transmitted speech across real-time communication platforms [17]. Tsutsumi et al. evaluate phone calls, Telegram calls, and interviews, and show the value of mixing original and degraded speech for adaptation [18]. Accordingly, neither communication-domain evaluation nor in-domain fine-tuning is itself the novelty claimed here. RealComm contributes a paired hardware-oriented design that crosses directed device routes with acoustic or wired injection, caller operating mode, and a controllable denoising state. The design connects detection, quality measurements, and adaptation on the same sources and conditions.

C. Detector Representations

End-to-end detectors such as RawNet2 learn directly from waveforms [19]; AASIST combines spectral and temporal graph attention [20]. A different family builds on pretrained speech representations. Wav2vec 2.0 and HuBERT establish self-supervised representation learning [21], [22], with XLS-R extending its multilingual scale [23] and WavLM targeting a wider range of speech tasks [24]. Whisper learns through large-scale weak supervision [25], while the Seamless work provides the W2v-BERT 2.0 encoder used by Teffic-Audio [26].

These encoders require an authenticity classifier and appropriate training; a speech backbone alone is not a deepfake detector. Tak et al. combine a fine-tuned wav2vec 2.0 front end with augmentation for anti-spoofing [27], and XLSR-SLS selects useful self-supervised layers for detection [28]. Our comparison includes existing detectors built on these families and the released DF Arena model [29]. Its purpose is to examine complete systems under a shared paired evaluation, not to attribute every difference to architecture or parameter count.

D. Augmentation, Adaptation, and Quality Diagnostics

RawBoost introduces convolutive and additive waveform perturbations motivated in part by telephony [30]. Cohen et al. study compression and channel augmentation for voice anti-spoofing [31]. Neural audio compression, including EnCodec [32], adds another family of transformations, but differs from both conventional file encoding and the codecs used to approximate call transmission. Together with the real-channel adaptation results in [18], this literature motivates comparing simulated variation, actual call recordings, and their combination under a common initialization.

RealComm-Aug organizes established operations by their position in a plausible signal path. We evaluate the resulting recipe on real recordings; we do not claim that its individual filters, noise sources, or codec operations are new. Quality diagnostics provide a separate view: Seed-TTS evaluation measures content accuracy and prompt-referenced speaker similarity [1], [33], while DNSMOS P.835 predicts perceptual speech quality [34]. Our paired analysis asks whether changes in such measures accompany detector degradation, without assuming that intelligibility, voice similarity, or perceived quality measures authenticity.

III. The RealComm Corpus

A. Call Acquisition and Conditions

Figure 1 summarizes the acquisition chain. Digital source speech is injected into a caller device, which initiates the call. The receiver records the received signal internally. This setup produces a digital reference and multiple call recordings for each source utterance.

1) Input configuration: Over-the-air (OtA) injection plays a source waveform through an external loudspeaker, allowing it to propagate through air into the caller's microphone. We record both handset and speakerphone operation of the caller. Line-to-microphone (LtM) injection connects the waveform to the microphone input through a cable, bypassing the additional air-propagation segment. The resulting configurations are OtA–Handset, OtA–Speakerphone, and LtM. "Handset" denotes the call operating mode; it is not an additional source or a different test partition.

2) Directed device routes: The training/development ring is Mate70 Pro–iPhone–Samsung–Oppo–Mate70 Pro. The test ring is Mate70 Pro–Samsung–iPhone–Vivo–Mate70 Pro. Adjacent devices communicate in both directions, giving eight directed routes per ring. Compared with all twelve directed pairs among four devices, the ring reduces acquisition requirements while retaining two neighbors per device in both roles. Its partial overlap permits comparisons between a shared route, an unseen pairing of covered devices, and an unseen device in either role.

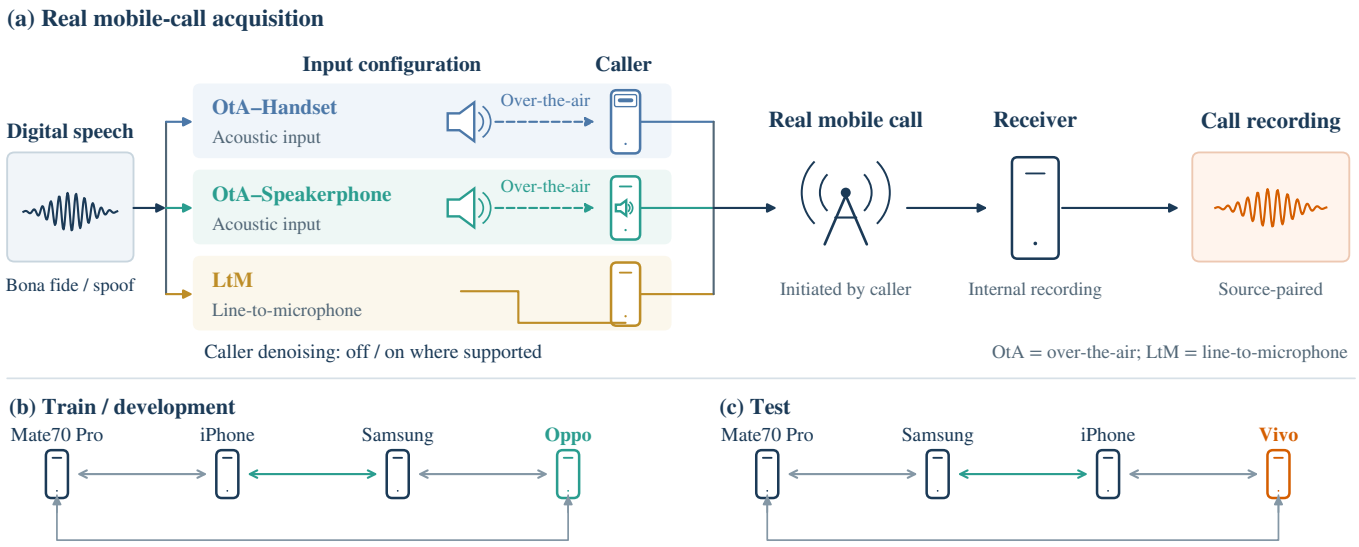

Fig. 1. RealComm acquisition and device partitions. The upper panel shows acoustic and wired injection into the caller, followed by a real mobile-call link and receiver-side recording. The lower panels show the bidirectional training/development and test rings; the return edge closes each ring. The Samsung–iPhone pair is shared; Oppo appears only in training/development and Vivo only in testing. Call modes and denoising settings refer to the caller.

3) User-controlled denoising: Eight routes and three input configurations give 24 base conditions. For configurations supporting a user-controlled denoising switch, we additionally record the on state. This adds 20 training-side and 18 test-side conditions, for totals of 44 and 42. We retain three distinct labels: *off*, *on*, and *unavailable*. The last means that no controllable switch exists in that configuration; it is not treated as off. Conversely, off means only that the exposed function is disabled, not that all internal speech processing has been removed.

B. Source Speech and Partitioning

The digital pool contains 2,800 utterances, balanced between bona fide and synthetic speech and between English and Mandarin. Bona fide speech, target text, and reference prompts originate from Seed-TTS Eval [1], [33]. Synthetic speech is generated by CosyVoice2 [2], F5-TTS [3], FlexiVoice [35], IndexTTS2 [36], MaskGCT [37], Vevo2 [38], and Minimax [39], with 200 utterances per system. Source selection uses content-accuracy and audio-quality screening; detector scores are not used to select samples. Consequently, the source pool is a quality-screened cohort rather than an unrestricted sample of all generation outcomes.

We group related content and reference-audio identities before partitioning. An initial 8:2 allocation reserves 560 digital sources for testing and 2,240 for the training side. A second 8:2 allocation divides the latter into 1,792 train and 448 development sources. All digital and call versions of a source inherit the same partition. This prevents a recorded version of an utterance from crossing into a different split from its digital counterpart.

The acquisition subset contains 224 train and 56 development sources, each recorded under 44 conditions. All 560 test sources are recorded under 42 conditions. Table I distinguishes independent source counts from recording counts. RealComm-Train comprises the train and development recordings, while RealCommBench comprises the test recordings. The smaller source set on the training side controls the cost of repeated real acquisition; the complete digital training pool remains available for other protocols.

TABLE I
SOURCE-LEVEL PARTITIONS AND REAL-CALL RECORDING COUNTS.

Split	Digital sources	Recorded sources	Conditions per source	Call recordings
Train	1,792	224	44	9,856
Dev	448	56	44	2,464
Test	560	560	42	23,520

The train and test sets cover the same seven generation systems. The intended generalization axes are held-out source content and communication paths, not unseen generators. Multiple recordings of a source expand condition coverage but do not add independent content. The balanced source and acquisition design also ensures that neither authenticity class is uniquely associated with a recording condition.

C. One Test Set, Several Analysis Dimensions

All 23,520 call recordings form a single test set. We report its pooled EER as the principal call-domain result. Input configuration, denoising state, and directed route are annotations for analyzing that same set, rather than alternative test protocols. OtA–Handset, OtA–Speakerphone, and LtM contain 7,840, 8,960, and 6,720 recordings, respectively. Their different sizes reflect the availability of denoising switches.

For adaptation analysis, we summarize routes in four groups relative to the RealCommTrain ring: Samsung↔iPhone (seen route), Mate70 Pro↔Samsung (unseen pairing), and Vivo as caller or receiver. “Seen” refers to the real-training acquisition

design; it does not mean that the zero-shot models have encountered those calls. Group differences may also reflect the mixture of input and denoising conditions, so they are not interpreted as isolated hardware effects.

IV. Experimental Protocol

A. Models and Waveform Input

The zero-shot evaluation uses WavLM-L (Teffic), Whisper (Teffic), AASIST (Teffic), Teffic-Audio, XLSR-SLS, and DF Arena 500M V1. We shorten the first three names to WavLM-L, Whisper, and AASIST, and the last to DF Arena. Teffic-Audio uses a W2v-BERT 2.0 encoder with an utterance-level pooling and classification head. These are existing trained detectors, not speech encoders used without a detection head. The existing-weight condition is denoted F and performs no RealComm parameter updates. Subsequent adaptation experiments use WavLM-L, AASIST, and Teffic-Audio.

Each model evaluates all 24,080 files: 560 digital test sources and 23,520 paired call recordings. Audio is converted to mono at 16 kHz. WavLM-L, Whisper, and Teffic-Audio read at most the first 64,600 samples; shorter utterances retain their variable length. AASIST, XLSR-SLS, and DF Arena first read the leading 64,000 samples and then cyclically pad to 64,600 samples as required by their interfaces. Each file receives one forward pass, with no sliding-window averaging. These model-specific rules are used consistently within each digital/call comparison. Cross-model differences therefore include both the trained system and its input interface.

We save one scalar score and one embedding from the same inference pass per file. Scores are oriented so that larger values indicate synthetic speech. Embeddings retain the recorded extraction layer and pooling convention without additional normalization. The present main analysis uses detection scores; embeddings support subsequent within-model diagnostics and are not used to select test samples.

B. Metrics and Matched Comparisons

We report EER in percent. Let $s_m(x)$ be the spoof-oriented score of model m on waveform x. For digital sources $\mathcal{D}$ and call recordings $\mathcal{C}$, the domain gap is

$$\Delta_m = \mathrm{EER}_m(\mathcal{C}) - \mathrm{EER}_m(\mathcal{D}), \tag{1}$$

expressed in percentage points. The pooled call EER is computed directly from all scores in $\mathcal{C}$, not by averaging subgroup EERs. A per-generator EER uses that generator's spoof samples and the same-domain bona fide reference set.

Input-path comparisons are supplemented by matched analyses using the same sources, routes, and shared denoising labels. Denoising off/on comparisons use 10,080 pairs with identical source, directed route, and input configuration. The 3,360 unavailable recordings contribute to the full test result but not to switch comparisons. When estimating uncertainty in domain differences, we use 2,000 paired bootstrap replicates over 148 source-association clusters, retaining all digital and call versions in each sampled cluster. These intervals characterize source uncertainty under the fixed acquisition conditions, not variability across new devices or independent recording campaigns.

All adaptation results are single-run checkpoint results. We preserve full-precision metrics for differences and round displayed EERs to two decimals. Thus a difference calculated from full-precision scores can differ by 0.01 points from subtracting two displayed values.

TABLE II
Zero-shot EER (%). The difference is in percentage points. Digital and call utterances are source-paired.

Model	Digital	Call	ΔEER
WavLM-L	4.64	40.00	+35.36
Whisper	15.71	47.88	+32.17
AASIST	7.50	40.19	+32.69
Teffic-Audio	0.00	37.95	+37.95
XLSR-SLS	38.57	50.31	+11.73
DF Arena	7.86	38.77	+30.91

V. Zero-Shot Detection Results

A. A Large Digital-to-Call Gap

Table II shows higher EER after call transmission for every detector. The increase ranges from 11.73 to 37.95 points; five of the six models deteriorate by more than 30 points. WavLM-L rises from 4.64% to 40.00%, AASIST from 7.50% to 40.19%, and DF Arena from 7.86% to 38.77%. Teffic-Audio records no errors at its digital EER operating point but reaches 37.95% on the call set. Since the source speech is paired, this large discrepancy cannot be attributed simply to using a different set of generated utterances at test time.

XLSR-SLS has a smaller absolute increase, from 38.57% to 50.31%. Its digital performance is already weak on this cohort, so the smaller gap is not evidence of superior channel robustness. This distinction matters when interpreting robustness scores: a low change from an unfavorable baseline is different from preserving effective discrimination. We retain all six models and their original score orientations rather than reversing or excluding a model after inspecting its test EER.

B. Acoustic and Wired Injection

Every model performs better on LtM than on either OtA configuration in Fig. 2(a). For WavLM-L, EER is 23.72% on LtM, compared with 41.96% on handset and 45.33% on speakerphone. Teffic-Audio exhibits the same ordering, with 16.46%, 41.10%, and 41.70%, respectively. The advantage of LtM also persists in comparisons matched for shared routes and denoising labels.

These observations associate the additional external-playback and microphone-capture path with substantially greater detection difficulty. They are consistent with better preservation of useful signal structure when the extra acoustic segment is bypassed. However, the recordings do not separately manipulate loudspeaker response, room propagation, microphone response, and internal processing; they cannot establish which component removes a particular spoofing artifact. Handset and speakerphone also have no common ordering across all detectors.

Fig. 2. Zero-shot EER by (a) input configuration and (b) denoising state. Models are shown on the horizontal axis and EER on the vertical axis. Input groups retain their observed condition composition; off/on use matched subsets of 10,080 recordings each. Lower is better.

C. Denoising, Language, and Generation Systems

The pooled switch effect in Fig. 2(b) is smaller and model-dependent. Denoising on reduces EER for WavLM-L, Whisper, AASIST, Teffic-Audio, and DF Arena, while increasing it for XLSR-SLS. AASIST changes from 42.70% to 40.18% on the paired off/on subsets. This predominantly favorable pooled direction does not imply a benefit for every recording condition or every detector.

Detailed language and generator results are provided in the supplementary tables. Both English and Mandarin show increased EER after transmission for all six models. Their relative difficulty is not consistent across systems: AASIST has lower call EER in English, whereas DF Arena has lower call EER in Mandarin. Each language contains 280 digital sources and 11,760 call recordings. Differences between these cohorts involve their speakers and content as well as language, and are not evidence that one language is intrinsically easier to detect.

Five models deteriorate for all seven generation systems. XLSR-SLS improves on CosyVoice2 and Minimax relative to its digital results, but its EER remains high in both domains. The per-generator comparisons establish breadth within the sampled generator set; they do not test unseen-generator generalization. Together, these results motivate examining properties of the transmitted signal rather than explaining failures solely by a particular language or synthesizer.

VI. Speech Quality, Spectral Changes, and Detection

A. Paired Speech-Quality Evaluation

Content accuracy and speaker similarity follow Seed-TTS Eval [33]. English WER is computed with Whisper-large-v3 and Mandarin CER with Paraformer-zh [40]. Both the digital and call waveform are evaluated against the same target text. Reported WER/CER values are means of per-utterance error rates, rather than error counts pooled over all words or characters.

Speaker similarity (SIM) uses the official WavLM-large speaker-verification model. For a reference prompt p_i and source x_i, we compare $\cos(e(p_i), e(x_i))$ with $\cos(e(p_i), e(x_{i,c}))$, where $x_{i,c}$ is the call recording under condition c. The reference prompt is held fixed across domains. SIM is reported separately for bona fide and synthetic speech; it measures similarity to that prompt, not waveform fidelity to the digital source. The evaluators are separate models from the authenticity detectors, although they can share a backbone family.

Overall quality is measured by DNSMOS P.835 OVRL [34]. This is a non-intrusive predictor of overall speech quality, evaluated independently on each waveform. Digital speech supplies the paired baseline for its change; it is not an input reference to DNSMOS itself. We use these automatic metrics to characterize the recordings, not as substitutes for subjective listening or security performance.

Transmission raises mean English WER from 0.71% to 10.12% and Mandarin CER from 1.19% to 7.18%. Bona fide SIM falls from 0.747 to 0.503, synthetic SIM from 0.754 to 0.537, and DNSMOS from 3.32 to 2.57. Among input paths, LtM has the lowest recognition error and the highest SIM and DNSMOS. For example, its DNSMOS is 3.03, compared with 2.36 and 2.42 for handset and speakerphone. The ordering of input-path quality therefore resembles the detection ordering in Section V.

Denoising behaves differently across configurations even after restricting the comparison to four common routes. On speakerphone, enabling it raises English WER and Mandarin CER by 3.96 and 3.48 points. On LtM, the corresponding means fall by 2.72 and 1.38 points. Both SIM categories likewise decrease on speakerphone and increase on LtM. Pooling the configurations can consequently obscure opposing effects. These measurements describe the selected routes and

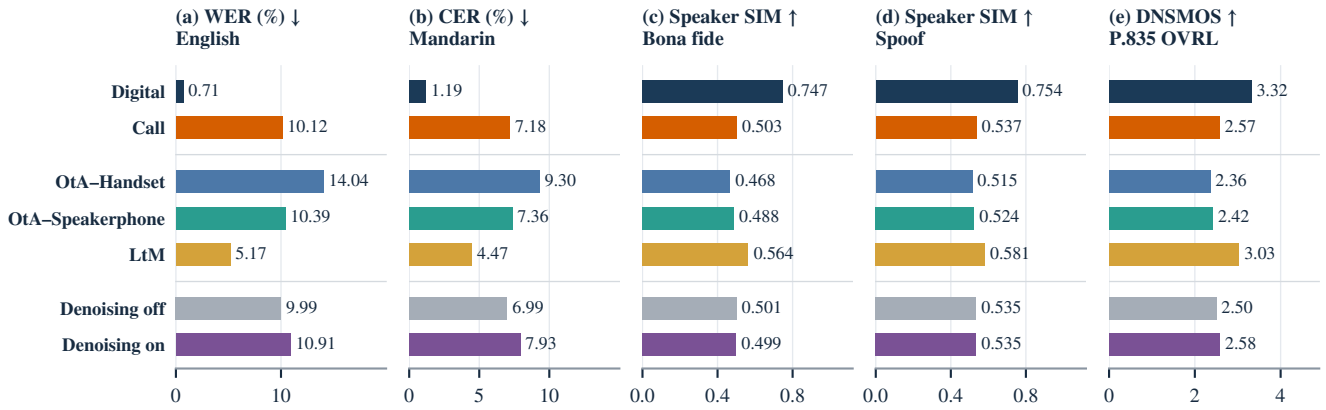

Fig. 3. Mean quality metrics in the digital and call domains, by input configuration, and for matched denoising off/on subsets. Each metric retains its own scale. WER and CER are averaged over utterances in the corresponding language. SIM is prompt-referenced and separated by authenticity class. DNSMOS denotes P.835 OVRL. Unavailable denoising cases remain in the pooled call result but are omitted from the switch panels.

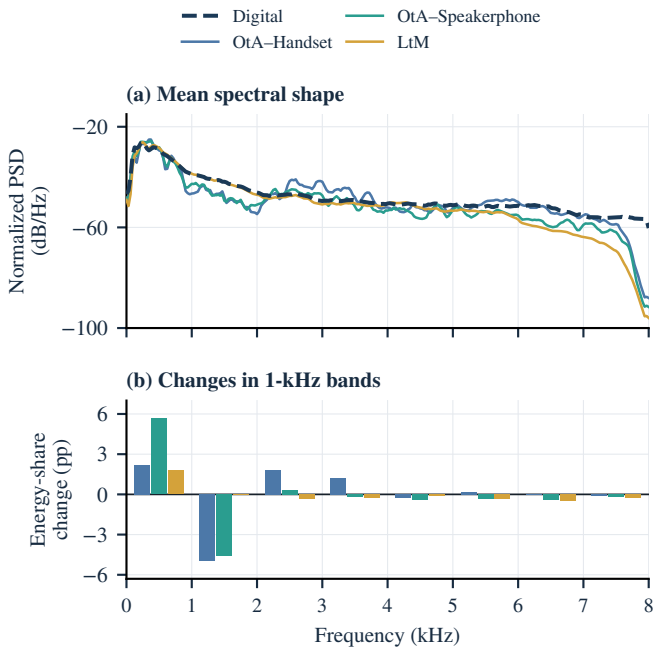

Fig. 4. Paired spectral changes. Top: normalized mean power spectral density (PSD). Bottom: changes in the energy share of successive 1-kHz bands, in percentage points relative to digital speech. The panels share a 0–8 kHz axis. The digital curve is dark blue and dashed.

should not be interpreted as a universal recommendation to enable or disable denoising.

B. Frequency-Dependent Changes

We estimate the PSD of each utterance using a common analysis procedure at 16 kHz and normalize it by its total power. For each input configuration, call PSDs are first averaged within source, then equally across the 560 sources. Averaging occurs in the linear domain before conversion to decibels. This procedure describes changes in spectral shape while preventing a source with more recordings from receiving greater weight.

Figure 4 shows pronounced attenuation near 7.5–8 kHz for all three call configurations. The 7–8 kHz energy share decreases from 0.248% in digital speech to 0.123%, 0.058%, and 0.015% for handset, speakerphone, and LtM. The changes also involve lower frequencies: both OtA configurations substantially reduce the 1–2 kHz share, whereas LtM remains close to the digital reference in that band.

Two distinctions are important. First, normalized energy shares do not measure absolute gain; increasing one band's share can result from attenuation elsewhere. Second, a decline near the 8-kHz analysis limit does not identify the negotiated call codec or its cutoff frequency. Resampling and device filters can also contribute. Indeed, LtM has the smallest 7–8 kHz share while yielding the best detection performance, already suggesting that preservation of this high-frequency tail alone cannot explain the input-path ranking.

C. Condition-Level Association with Detection Degradation

Each of the 42 conditions contains the same 560 test sources. For a quality measure q, we compute a condition-level mean deterioration

$$d_{q,c} = \frac{1}{N_q} \sum_{i \in \mathcal{I}_q} \epsilon_q \big[q(x_{i,c}) - q(x_i) \big], \tag{2}$$

where $\epsilon_q = +1$ for WER/CER and -1 for SIM/DNSMOS. $\mathcal{I}_q$ contains the applicable sources: the corresponding language for WER/CER and all sources for the other metrics. We correlate $d_{q,c}$ across conditions with

$$\Delta_{m,c} = \mathrm{EER}_m(\mathcal{C}_c) - \mathrm{EER}_m(\mathcal{D}) \tag{3}$$

using Spearman's rank coefficient. WER/CER use language-matched EERs. A positive coefficient means that conditions with larger quality deterioration tend to exhibit greater detection deterioration. For spectral analysis, we substitute the

Fig. 5. Spearman association across 42 recording conditions between paired quality or spectral changes and EER increases. WER and CER use language-matched EERs; SIM pools both authenticity classes equally. The high-frequency row describes a reduction in relative energy, not necessarily worse perceived quality.

decrease in normalized 4–8 kHz energy share; this is a spectral change, not intrinsically a quality score.

Except for XLSR-SLS, all models exhibit positive associations for the four quality measures in Fig. 5. DNSMOS deterioration has a correlation of approximately 0.72–0.90 with detection degradation for those five models, compared with 0.17 for XLSR-SLS. Thus, poorer-quality call conditions tend to be more difficult to classify, but the strength and consistency of this relationship depend on the detector.

The relationship is not explained entirely by pooling OtA and LtM: within LtM, DNSMOS deterioration remains positively associated with EER increases for all six models. It nevertheless does not predict every intervention. On the four common handset routes, enabling denoising improves mean recognition error, both SIM measures, and DNSMOS, while XLSR-SLS EER rises by 3.66 points. Better automatic quality scores therefore do not guarantee a better authenticity decision.

The 4–8 kHz energy-share decrease shows weaker and mixed associations, from -0.22 to 0.15. High-frequency changes are visible, yet their magnitude alone does not rank condition difficulty reliably. This finding favors joint examination of acquisition and processing conditions over a single bandwidth explanation. These correlations are descriptive: conditions share sources and devices, and both quality and detection can respond to the same processing chain. Quality metrics also use complete utterances, while detection uses the prefix protocol of Section IV.

VII. Simulated Augmentation and Real-Call Adaptation

A. RealComm-Aug

RealComm-Aug applies stochastic waveform transformations in the order suggested by physical acquisition and communication. For an input (x, y), a scenario z and its parameters θ are sampled, yielding

$$\widetilde{x} = T_{z,\theta}(x), \qquad z \sim \text{Categorical}(\boldsymbol{\pi}), \quad \theta \sim p(\theta \mid z). \quad (4)$$

The same sampling distribution is used for bona fide and synthetic speech, and the authenticity label y is preserved. Figure 6 separates acoustic environment, device response and processing, file coding, and call transmission. Intermediate exits incorporate conventional noise, reverberation, device, and file perturbations without requiring a simulated call on every example.

1) Acoustic environment: Room impulse responses from RIRS_NOISES and Graphi07 simulate reverberation, while MUSAN [41] and Gaussian noise supply background disturbances. Reverberation and background noise can operate independently or together. Environmental background sound is introduced before downstream device and communication processing, so it is transformed together with the speech. The direct, LtM-like call path bypasses this additional air-propagation stage.

2) Device response and processing: Device impulse responses from micirp, MADIR, and POLIPHONE, or a smooth equalization curve, perturb the frequency response. Capture noise is distinct from environmental background sound. Low-amplitude gating, soft clipping or convolutive nonlinear perturbation, level control, dynamic compression, and linear bit-depth quantization provide amplitude and nonlinear variability. On the direct call path, optional equalization is used without an additional acoustic microphone response. These operations approximate families of device effects; they do not reproduce a particular RealComm phone model.

3) File and communication coding: MP3, AAC, and Vorbis form the file-coding branch. The call branch combines frequency filtering, round-trip resampling, and candidates drawn from GSM, Opus, Speex, G.711, and G.723.1. File and communication coding are sampled as different output scenarios rather than sequentially imposed on every example. The codec candidates create training distortions and are not claims about the codecs negotiated in the real recordings.

4) Packet-loss simulation and receiver output: Independent or burst masks operate on 20-ms decoded waveform frames, replacing masked segments with silence or a repeated preceding frame. This approximates audible consequences of packet loss; it does not discard actual encoded packets or reproduce a complete network stack. The received waveform is then passed directly to the detector. There is no receiver-side loudspeaker replay, room convolution, or added environmental noise.

B. Training Conditions

Table III defines the comparisons. A adapts the detector using simulated perturbations of its original digital training data. T adds the 9,856 RealCommTrain training recordings while retaining the original digital augmentation. A+T combines the same real-call data with RealComm-Aug on digital samples. Development recordings are kept out of training. We do not additionally insert the matched digital source subset from DigitalPool into these experiments. A+T is initialized from F, rather than obtained by continuing an already adapted A or T model.

The model architecture and model-specific training recipe are maintained within each comparison. Existing model-internal neural-codec augmentation is retained as part of that

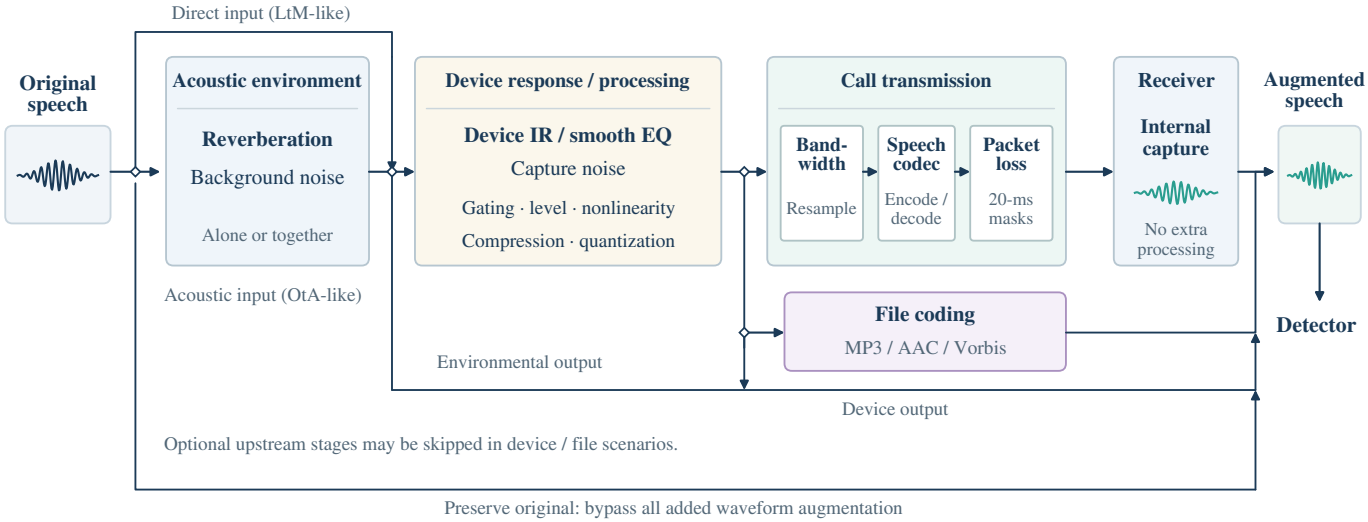

Fig. 6. RealComm-Aug processing order and output scenarios. One scenario determines the stages and output exit; individual operations are sampled within that scenario. The upper bypass represents direct, LtM-like input, while the acoustic path passes through the environmental stage. Environmental, device, and file outputs provide non-call augmentation. Upstream stages may be skipped in device/file scenarios. Call output is captured internally without receiver-side replay. IR: impulse response; EQ: equalization.

TABLE III
ADAPTATION SETTINGS. "ORIGINAL DIGITAL DATA" DENOTES THE DETECTOR'S EXISTING TRAINING MIXTURE, NOT THE DIGITAL COUNTERPARTS OF REALCOMMTRAIN. EVERY ADAPTED MODEL STARTS FROM ITS CORRESPONDING F CHECKPOINT.

Setting	Training data	Data-side waveform processing
F	No parameter updates	Evaluation preprocessing only
A	Original digital data	RealComm-Aug
T	Original digital data + RealCommTrain/train	Original augmentation on digital examples; no additional data-side augmentation on call recordings
A+T	Same data as T	RealComm-Aug on digital examples; no additional data-side augmentation on call recordings

recipe; EnCodec [32] is not a newly added RealComm-Aug stage. Data-side bypass for real recordings therefore describes only the new waveform pipeline, not a claim that every operation internal to the training model is disabled. No augmentation is applied at test time.

The documented staged recipe retains original waveforms with probability 0.20 and allocates 0.08, 0.06, and 0.06 to acoustic-only, device-output, and file-output scenarios. Direct-call and acoustic-call scenarios each have probability 0.30. Operation parameters are sampled within the chosen scenario, so the route probability is different from the marginal probability of any individual operation. This is a fixed recipe, not the outcome of a reported component-wise optimization study. The model-specific checkpoints and their provenance are retained with the experimental results.

VIII. ADAPTATION RESULTS AND ANALYSIS

A. Overall Adaptation and Digital-Domain Retention

Table IV shows that A lowers call EER by 1.29, 1.69, and 3.02 points for WavLM-L, AASIST, and Teffic-Audio, respectively. T produces larger reductions of 4.52, 2.23, and 14.11 points. Teffic-Audio benefits most from real-call training, moving from 37.94% to 23.84%.[1]

TABLE IV
ADAPTATION EER (%). BOLD MARKS THE BEST SETTING WITHIN EACH MODEL AND DOMAIN. ALL ADAPTED SYSTEMS START FROM F.

Model	Domain	F	A	T	A+T
WavLM-L	Digital	4.64	4.29	**3.21**	**3.21**
	Call	40.00	38.71	35.48	**34.16**
AASIST	Digital	7.50	8.93	**6.07**	8.21
	Call	40.19	38.49	37.96	**36.90**
Teffic-Audio	Digital	**0.00**	0.36	0.36	**0.00**
	Call	37.94	34.92	23.84	**22.37**

Combining augmentation and real recordings yields the lowest pooled call EER for every model: 34.16% for WavLM-L, 36.90% for AASIST, and 22.37% for Teffic-Audio. Relative to T, the additional improvements are 1.33, 1.06, and 1.46 points. These consistent overall gains support retaining simulated digital-domain variation when real-call data are available. They do not show that simulation can replace real acquisition, since A alone remains worse than T for all three systems.

Digital-domain performance does not necessarily improve in parallel. WavLM-L has 3.21% digital EER under both T and

[1]The adaptation comparisons use a matched FP32 re-evaluation of the same Teffic-Audio F checkpoint. Its call EER is 37.94%, compared with 37.95% in the frozen six-model submission of Table II. Both results are retained with their inference metadata.

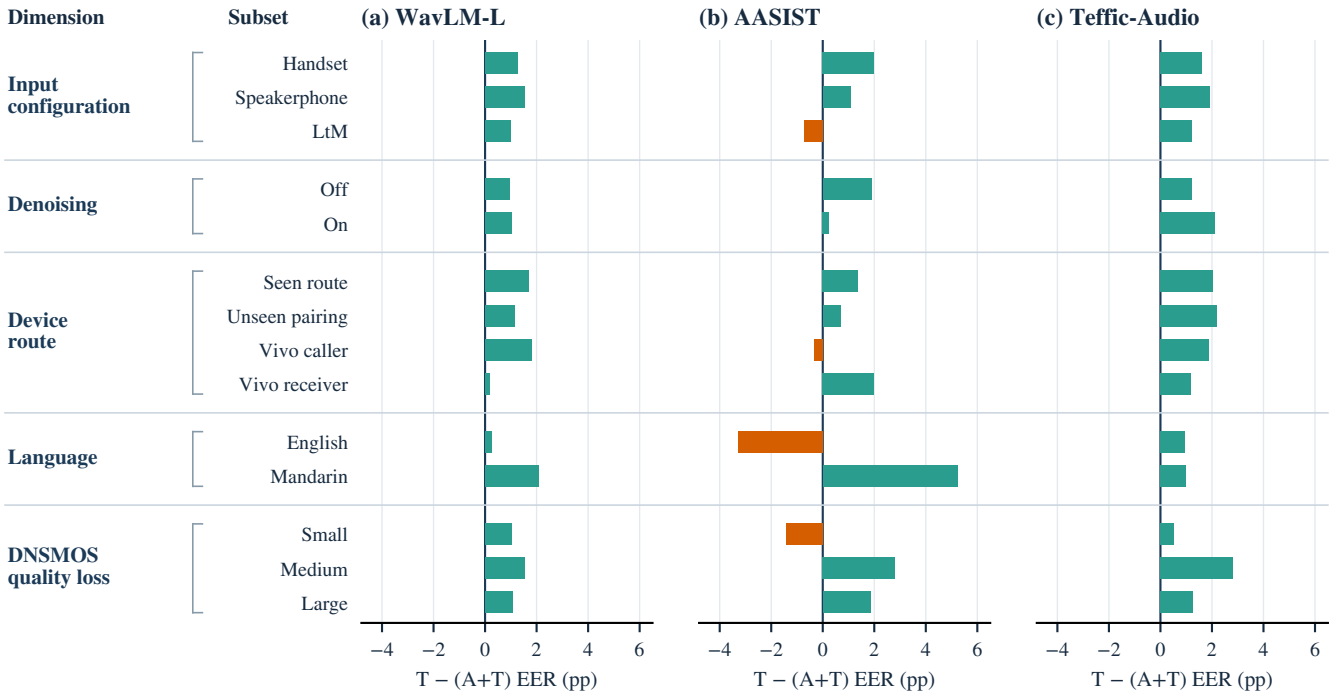

Fig. 7. Incremental benefit of combined adaptation over real-call training: T minus A+T EER, in percentage points. Positive teal bars denote improvement; negative orange bars denote regression. Labeled blocks group the subsets by input configuration, denoising state, device route, language, and DNSMOS quality-loss stratum. All groups are defined independently of detection scores. Full F/A/T/A+T values appear in the supplementary tables.

A+T, compared with 4.64% under F. Teffic-Audio remains at or below 0.36% across all settings. AASIST changes from 6.07% under T to 8.21% under A+T. Its call-domain gain therefore comes with a measurable digital-domain tradeoff. Reporting only the call result would conceal this behavior.

B. Input Configuration and Denoising

Figure 7 summarizes the incremental gains. For WavLM-L, A+T improves on T by 1.25 points on handset, 1.54 on speakerphone, and 0.98 on LtM. Teffic-Audio also improves in all three configurations. AASIST instead improves on the two OtA subsets while LtM worsens from 23.57% to 24.29%. The benefit of combination is therefore broad but not uniform.

Acoustic injection remains the principal difficulty after adaptation. For every model and setting, LtM EER is lower than both OtA EERs. Even Teffic-Audio A+T, the strongest overall system in this comparison, reaches 27.86% on handset and 28.26% on speakerphone, versus 7.20% on LtM. The remaining gap indicates that the current combination has not resolved the effects associated with external playback and acoustic capture.

A+T improves on T in both denoising states for all three models, so its pooled benefit is not confined to a single switch setting. However, the direction of the denoising effect can change after adaptation. Teffic-Audio F has lower EER with denoising on, whereas T has lower EER with it off. A+T nearly equalizes the two, at 23.63% and 23.73%. Robustness to processing states and an average benefit from activating a particular state are distinct properties.

C. Held-Out Device, New Routes, and Languages

T and A+T improve on F in all four route groups for each model. WavLM-L and Teffic-Audio additionally improve from T to A+T in every group. AASIST improves in three groups but changes from 37.54% to 37.86% when Vivo is the caller. The results demonstrate transfer beyond the route shared with training, including the held-out Vivo device. Their scope is this particular device ring; they do not establish performance on arbitrary new hardware or communication services.

WavLM-L and Teffic-Audio also benefit from A+T over T in both language cohorts. For AASIST, Mandarin EER falls from 42.47% to 37.23%, while English rises from 33.20% to 36.48%. Thus, although the adaptation benefit is not restricted to one language across the study, AASIST exhibits a language-dependent tradeoff that its pooled result obscures.

D. Quality Strata and Individual Conditions

We reuse the quality measurements from Section VI to divide conditions into small, medium, and large mean DNSMOS-decrease groups, each containing 14 conditions and 7,840 recordings. The same score-independent grouping is used for every model and setting (supplementary tables). WavLM-L and Teffic-Audio improve from T to A+T in all three strata. AASIST improves in the medium and large strata but worsens in the small-decrease group.

Conditions with greater quality loss remain more difficult even after adaptation: all three models under all four settings have increasing EER across the three strata. Teffic-Audio A+T

reaches 10.97% in the small-decrease group and 31.20% in the large-decrease group. This stratification summarizes a mixture of path, route, and denoising effects, rather than identifying an independent causal effect of perceived quality.

At the finer level of the 42 individual recording conditions, A+T improves on T in 25 conditions for WavLM-L, 27 for AASIST, and 28 for Teffic-Audio. It worsens in 12, 11, and 6 conditions, respectively; the remaining conditions tie. The best pooled EER therefore coexists with local regressions, including for the two pretrained-encoder systems.

E. Model-Dependent Response to Augmentation

WavLM-L and Teffic-Audio have a consistent broad-group trend in Fig. 7: A is better than or tied with F in every displayed input, denoising, route, language, and quality subgroup, and A+T is better than T. AASIST also improves overall, but some group gains are accompanied by losses elsewhere, including English and LtM under A+T.

Model capacity and pretrained representations provide a plausible joint explanation. WavLM-L and Teffic-Audio use large pretrained speech backbones, while AASIST has a compact task-specific architecture. Richer initial representations and greater capacity may help accommodate channel variation while retaining authenticity cues. This interpretation is consistent with prior corruption analyses [15], but the present comparison does not isolate these factors, since architecture and pretraining change together with capacity. It therefore supports a model-dependent response, not a demonstrated capacity bottleneck. Moreover, AASIST improves more individual conditions than WavLM-L in the T-to-A+T comparison, so broad-group consistency should not be generalized into a universal ranking of robustness.

IX. Discussion and Conclusion

RealComm exposes a gap between detecting a digital source file and detecting its received mobile-call version. The paired design connects that gap to acquisition conditions without replacing the source content. Across six existing detectors, acoustic injection is consistently more difficult than wired injection. Quality deterioration tracks condition-level detection degradation for most models, whereas the change in high-frequency energy share alone has limited explanatory value. Neither an improved quality score nor an apparently preserved bandwidth should therefore substitute for evaluation of received-call authenticity.

The adaptation experiments show complementary practical benefits. Simulated augmentation improves the three evaluated models; real-call training provides larger gains; combining the two yields the lowest pooled call EER for each model. Gains extend to unseen route combinations and the held-out device in the current design. Remaining errors are concentrated in challenging acoustic conditions, and AASIST illustrates that pooled improvement can coexist with digital-domain or subgroup regressions.

Several boundaries delimit these findings. The corpus uses a limited hardware set, two languages, and a fixed set of seven generators shared across splits. The denoising label represents an exposed device control rather than an isolated signal-processing algorithm. The analyses are descriptive, and repeated versions of a source are not independent content samples. Adaptation results are single runs; a frozen-to-adapted comparison also includes additional optimization, while the more focused T versus A+T comparison changes the digital augmentation recipe. Controlled component ablations, repeated seeds, real-data scaling experiments, and external communication-domain evaluation are needed to identify why the recipe works and how far its benefit transfers.

Within these boundaries, the results support evaluating digital references and real-call recordings together, retaining condition labels, and combining simulated digital variation with actual call examples. RealComm provides the paired data and analysis structure needed to study these questions at the signal, detector, and adaptation levels.

References

[1] P. Anastassiou, J. Chen, J. Chen, Y. Chen *et al.*, "Seed-TTS: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024. [Online]. Available: https://arxiv.org/abs/2406.02430

[2] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "CosyVoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024. [Online]. Available: https://arxiv.org/abs/2412.10117

[3] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024. [Online]. Available: https://arxiv.org/abs/2410.06885

[4] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Interspeech 2019*, 2019, pp. 1008–1012.

[5] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen *et al.*, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[6] X. Wang, H. Delgado, H. Tak, J. weon Jung, H. jin Shim, M. Todisco *et al.*, "ASVspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech," *Computer Speech & Language*, vol. 95, no. 101825, p. 101825, 2026.

[7] X. Wang, H. Delgado, N. Evans, X. Liu, T. Kinnunen, H. Tak *et al.*, "ASVspoof 5: Evaluation of Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech," *arXiv preprint arXiv:2601.03944*, 2026. [Online]. Available: https://arxiv.org/abs/2601.03944

[8] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," in *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021. [Online]. Available: https://arxiv.org/abs/2111.02813

[9] N. M. Müller, P. Kawa, W. H. Choong, E. Casanova, E. Gölge, T. Müller *et al.*, "MLAAD: The Multi-Language Audio Anti-Spoofing Dataset," in *International Joint Conference on Neural Networks (IJCNN)*, 2024. [Online]. Available: https://arxiv.org/abs/2401.09512

[10] Y. Lu, Y. Xie, R. Fu, Z. Wen, J. Tao, Z. Wang *et al.*, "Codecfake: An Initial Dataset for Detecting LLM-based Deepfake Audio," in *Interspeech 2024*, 2024, pp. 1390–1394. [Online]. Available: https://arxiv.org/abs/2406.08112

[11] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does Audio Deepfake Detection Generalize?" in *Interspeech 2022*, 2022, pp. 2783–2787.

[12] J. weon Jung, Y. Wu, X. Wang, J.-H. Kim, S. Maiti, Y. Matsunaga *et al.*, "SpoofCeleb: Speech Deepfake Detection and SASV in the Wild," *IEEE Open Journal of Signal Processing*, vol. 6, pp. 68–77, 2025.

[13] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio Deepfake Detection: A Survey," *arXiv preprint arXiv:2308.14970*, 2023. [Online]. Available: https://arxiv.org/abs/2308.14970

[14] Y. Zhang, G. Zhu, F. Jiang, and Z. Duan, "An Empirical Study on Channel Effects for Synthetic Voice Spoofing Countermeasure Systems," in *Interspeech 2021*, 2021, pp. 4309–4313.

[15] X. Li, P.-Y. Chen, and W. Wei, "Measuring the Robustness of Audio Deepfake Detection under Real-World Corruption," *arXiv preprint arXiv:2503.17577*, 2025. [Online]. Available: https://arxiv.org/abs/2503.17577

[16] H. Shi, X. Shi, S. Dogan, S. Alzubi, T. Huang, and Y. Zhang, "Benchmarking Audio Deepfake Detection Robustness in Real-World Communication Scenarios," in *2025 33rd European Signal Processing Conference (EUSIPCO)*, 2025, pp. 566–570.

[17] J. Xue, Z. Yi, Y. Huang, Y. Ren, Y. Chen, C. Fan, Z. Su, Y. Zhang, and B. Cai, "RTCFake: Speech deepfake detection in real-time communication," in *Findings of the Association for Computational Linguistics: ACL 2026*, 2026, pp. 5763–5775. [Online]. Available: https://aclanthology.org/2026.findings-acl.285/

[18] A. Tsutsumi, A. Gotoh, Y. Saito, H. Matsuura, and S. Shiota, "Domain Adaptation for Deepfake Audio Detection under Degraded Channel Conditions," in *Odyssey 2026*, 2026, pp. 9–14.

[19] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," *arXiv preprint arXiv:2011.01108*, 2020. [Online]. Available: https://arxiv.org/abs/2011.01108

[20] J. weon Jung, H.-S. Heo, H. Tak, H. jin Shim, J. S. Chung, B.-J. Lee *et al.*, "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6367–6371.

[21] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *arXiv preprint arXiv:2106.07447*, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[23] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal *et al.*, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," *arXiv preprint arXiv:2111.09296*, 2021. [Online]. Available: https://arxiv.org/abs/2111.09296

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022. [Online]. Available: https://arxiv.org/abs/2110.13900

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[26] Seamless Communication *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023. [Online]. Available: https://arxiv.org/abs/2312.05187

[27] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. Evans, "Automatic Speaker Verification Spoofing and Deepfake Detection Using Wav2vec 2.0 and Data Augmentation," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 112–119.

[28] Q. Zhang, S. Wen, and T. Hu, "Audio deepfake detection with self-supervised XLS-R and SLS classifier," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 6765–6773. [Online]. Available: https://github.com/QiShanZhang/SLSforASVspoof-2021-DF

[29] Speech Arena, "DF Arena 500M V1: Released antispoofing model," Model repository, 2026, accessed: Sep. 19, 2026. [Online]. Available: https://huggingface.co/Speech-Arena-2025/DF_Arena_500M_V_1

[30] H. Tak, M. Kamble, J. Patino, M. Todisco, and N. Evans, "Rawboost: A Raw Data Boosting and Augmentation Method Applied to Automatic Speaker Verification Anti-Spoofing," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6382–6386.

[31] A. Cohen, I. Rimon, E. Aflalo, and H. H. Permuter, "A study on data augmentation in voice anti-spoofing," *Speech Communication*, vol. 141, pp. 56–67, 2022.

[32] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://arxiv.org/abs/2210.13438

[33] ByteDance Speech, "Seed-TTS-Eval: Objective evaluation set and evaluation scripts," Software repository, accessed: Sep. 19, 2026. [Online]. Available: https://github.com/BytedanceSpeech/seed-tts-eval

[34] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. [Online]. Available: https://arxiv.org/abs/2110.01763

[35] D. Chen, X. Zhang, Y. Wang, K. Dai, L. Ma, and Z. Wu, "FlexiVoice: Enabling Flexible Style Control in Zero-Shot TTS with Natural Language Instructions," *arXiv preprint arXiv:2601.04656*, 2026. [Online]. Available: https://arxiv.org/abs/2601.04656

[36] S. Zhou, Y. Zhou, Y. He, X. Zhou, J. Wang, W. Deng *et al.*, "IndexTTS2: A Breakthrough in Emotionally Expressive and Duration-Controlled Auto-Regressive Zero-Shot Text-to-Speech," *arXiv preprint arXiv:2506.21619*, 2025. [Online]. Available: https://arxiv.org/abs/2506.21619

[37] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng *et al.*, "MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer," *arXiv preprint arXiv:2409.00750*, 2024. [Online]. Available: https://arxiv.org/abs/2409.00750

[38] X. Zhang, J. Zhang, Y. Wang, C. Wang, Y. Chen, D. Jia *et al.*, "Vevo2: A Unified and Controllable Framework for Speech and Singing Voice Generation," *arXiv preprint arXiv:2508.16332*, 2025. [Online]. Available: https://arxiv.org/abs/2508.16332

[39] B. Zhang, C. Guo, G. Yang, H. Yu, H. Zhang, H. Lei *et al.*, "MiniMax-Speech: Intrinsic Zero-Shot Text-to-Speech with a Learnable Speaker Encoder," *arXiv preprint arXiv:2505.07916*, 2025. [Online]. Available: https://arxiv.org/abs/2505.07916

[40] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition," *arXiv preprint arXiv:2206.08317*, 2022. [Online]. Available: https://arxiv.org/abs/2206.08317

[41] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015. [Online]. Available: https://arxiv.org/abs/1510.08484