DASF-Net: Dual-Adversarial Spectral Fusion for Cross-Lingual and Adversarially Robust Audio Deepfake Detection

Iftikhar Ahmed (2121019642), Meherun Nessa Mithila (2221649042)
Department of Electrical and Computer Engineering, *North South University*, Dhaka, Bangladesh
{iftikhar.ahmed, meherun.mithila}@northsouth.edu
CSE 445, Section 6 · Faculty: Mohammad Abdul Qayum · 6 September 2026

***Abstract*—Audio deepfake detectors are trained almost exclusively on high-resource languages and a narrow family of neural vocoders. Meeting a new language synthesised by a structurally different generator, they fail twice: the learned features encode language- and corpus-specific cues rather than synthesis artifacts, and the decision boundary stays trivially breakable by gradient-based perturbation. We propose DASF-Net, built on the observation that both failures are *adversarial* problems solvable by one formulation. It couples (i) a dual-branch front-end presenting each utterance as a sparse cepstral (LFCC) and a dense spectral (log-Mel) view to one Siamese-shared, ImageNet-pretrained Xception trunk; (ii) a Cross-Attention Spectral Fusion (CASF) module whose learned gate selects per utterance which view carries the artifact; and (iii) a dual-adversarial objective in which a gradient-reversed language discriminator drives language invariance while adaptive FGSM/PGD training drives perturbation robustness, trained by a three-stage curriculum. The design is instantiated on English WaveFake as source and Bengali BanglaFake as target — a pairing varying language and synthesis paradigm at once — under a Confound-Controlled Evaluation Protocol. This report covers design, architecture, algorithms and the experimental plan; the experiments have not been run and no results are reported.**

***Index Terms*—machine learning, audio deepfake detection, cross-lingual generalisation, adversarial robustness, domain-adversarial training, cross-attention fusion, Xception, Bengali speech, equal error rate**

I. Introduction, Objectives and Contributions

Neural vocoders and end-to-end TTS now produce speech listeners routinely fail to separate from genuine recordings [1]–[4], and the resulting misuse has driven a large body of detection work [5], [6]. That work is overwhelmingly English and built on one family of generators. Bengali, with over 230 million native speakers, has almost none: BanglaFake [7] first released Bengali deepfake audio at scale but trained no detector, and "Zero-Shot to Zero-Lies" [8] first benchmarked detectors on it, reporting 46.20% EER zero-shot and 24.35% EER fine-tuned (ResNet18).

Those numbers expose a gap that is not merely quantitative. A detector that transfers this poorly is not under-trained; it has learned the wrong thing — its features encode which corpus, speaker and language produced an utterance rather than what a synthesiser did to it. The same detectors are breakable by inaudible perturbations: Kawa *et al.* [15] degrade an LCNN from 0.022 to 0.785 EER under FGSM, and our group's prior study [9] reproduces this on a far larger Xception detector.

Motivation. Both failures share a structure: the network has latched onto a *nuisance direction* in feature space — language and corpus identity in one case, a locally exploitable gradient in the other. In both cases the standard remedy is an adversary that penalises exactly that direction. Domain-adversarial training [11] supplies the first; FGSM/PGD training [13], [14] the second. Placing both on one shared embedding is, to our knowledge, new for audio deepfake detection, and it turns the cross-lingual question from a measurement exercise into an architectural one.

Objectives. (1) Implement DASF-Net and verify correctness by shape and gradient tests. (2) Establish an in-language Bengali baseline and test whether the dense-over-sparse advantage reported for English CNNs [9] replicates on Bengali VITS audio, by EER against the 0.2435 EER ResNet18 baseline [8]. (3) Quantify zero-shot English↔Bengali transfer with and without the language adversary. (4) Measure whether adversarial robustness learned in one language survives transfer, on one unit-consistent ϵ grid.

Contributions. (1) DASF-Net, fusing sparse cepstral and dense spectral views over one Siamese-shared trunk via two-way cross-attention and a learned gate. (2) A dual-adversarial objective plus a three-stage curriculum making it trainable. (3) A Confound-Controlled Evaluation Protocol (CCEP). (4) A corrected, unit-consistent threat model resolving an inconsistency in our group's own earlier work. (5) An ablation-first design, so any gain can be attributed rather than merely observed.

II. Project Background

Context and rationale. Speech synthesis has moved in a decade from concatenative systems with audible seams to neural systems that are, for untrained listeners, transparent. The BanglaFake authors report a Mean Opinion Score of 3.40 for naturalness and 4.01 for intelligibility, with a t-SNE view of MFCC features showing real and synthetic clusters overlapping heavily [7]: the material already deceives, and simple cepstral features already fail to separate it. Detection research has not kept pace — detectors are validated within one language, and against passive forgeries rather than an

active gradient-computing adversary. A deployed countermeasure faces both conditions at once.

Historical and scientific context. Countermeasures grew out of ASVspoof [5], where the dominant design was LFCC with a GMM back-end; WaveFake [6] extended this to seven vocoders over LJSpeech [16] and found LFCC beating MFCC under a GMM. That ordering is back-end-specific: with a deep CNN, a dense 128-bin log-Mel spectrogram transfers far better into an ImageNet-pretrained Xception [10] than a sparse 20-coefficient representation, because it retains the 2-D time–frequency structure convolution and ImageNet initialisation both assume, giving 33–35% relative EER improvement [9]. In parallel, domain-adversarial networks [11] showed an encoder can be forced to be uninformative about domain while staying discriminative, and adversarial training [13], [14] became the standard evasion defence. This project joins these two lines.

Significance. Roughly seven thousand languages are spoken worldwide and detection resources exist for a handful. A method that transfers a detector into a language with no labelled deepfake data — using only unlabelled target audio and a language tag — is worth far more than one needing a new labelled corpus per language, and the design is not Bengali-specific.

III. Literature Review

Foundations. Four bodies of work underpin this project: benchmark datasets (ASVspoof [5], WaveFake [6], BanglaFake [7], built on SUST TTS [17] and Common Voice [18]); the LFCC-versus-Mel question and its back-end dependence [6], [9]; the gradient reversal layer [11], a minimax game between encoder and domain classifier implemented by a single backward-pass sign flip; and adversarial machine learning [13], [14], with attention [12] as the fusion mechanism used in CASF.

Gaps and limitations. (i) BanglaFake [7] trains no detector, so its detection difficulty was never established. (ii) Our prior work [9] is single-language and single-corpus, treats LFCC and log-Mel as mutually exclusive and lists hybrids as future work, downsamples to 16 kHz discarding the 8–11 kHz band, and reports FGSM and PGD at ϵ values two orders of magnitude apart, making the attacks non-comparable. (iii) "Zero-Shot to Zero-Lies" [8] measures the gap but does not close it, naming cross-lingual generalisation and adversarial robustness as open. (iv) An unflagged confound: BanglaFake's synthetic side is one male speaker while its real side spans seven speakers of both genders, so a detector can score highly by classifying gender; no published work controls for it. (v) No cross-lingual adversarial study exists.

Case studies. Kawa *et al.* [15] give the clearest case study of the adversarial failure, degrading an LCNN from 0.022 to 0.785 EER under FGSM and recovering most of it through adversarial training at moderate clean-accuracy cost. Our prior study [9] reproduces this at much larger scale, cutting FGSM error from 0.70 to 0.107 and PGD from 0.90 to 0.154, at the cost of clean EER rising from 0.034 to 0.088.

TABLE I
Dataset summary after High-Band-Preserving alignment

	WaveFake	**BanglaFake**
Role / language	source, English	target, Bengali
Real utterances	13,100	12,260
Synthetic utterances	104,885	13,260
Synthesis family	7, GAN / flow	1, VITS end-to-end
Speaker(s), synthetic	1 female	1 male
Speaker(s), real	1 female	7 (mixed gender)
Sample rate (retained)	22,050 Hz	22,050 Hz

Relation and advantage. This project repairs the limitations of two of these works: BanglaFake supplies target-domain data it never used for detection, and our prior Mel/Xception work supplies the source-domain method, which we extend rather than reuse. We contribute fusion instead of selection (taking up the hybrid direction [9] left open, with an interpretable gate as a by-product); closing the transfer gap rather than reporting it; preserving the discriminative 8–11 kHz band; a defensible evaluation under CCEP; and one ϵ grid for both attacks.

IV. Design Methodology and Proposed System

Overview. Both corpora are normalised to a common representation without discarding high-frequency content; every utterance is expressed as two complementary spectral views and encoded by a single shared backbone; the views are fused by a learned attention-based gate into one embedding carrying three heads; and that embedding is shaped by two adversaries at once under a three-stage curriculum (Fig. 1).

A. Datasets

WaveFake (English, source) [6] holds 117,985 utterances (13,100 real, 104,885 synthetic) from single-speaker LJSpeech [16], across seven GAN-based or GAN-adjacent flow vocoders that resynthesise a waveform from acoustic features of real speech. **BanglaFake** (Bengali, target) [7] holds 25,520 utterances: 12,260 real (10,000 SUST TTS Corpus, 2,260 Common Voice across five speakers) and 13,260 from one VITS model [4] trained on the SUST corpus. VITS generates speech directly from text with no intermediate acoustic stage, so it is structurally unlike every WaveFake generator, and because language and synthesis paradigm change together, transfer probes cross-lingual *and* cross-architecture generalisation at once (Table I).

B. Data Preprocessing Pipeline

Every utterance is brought to a common representation: mono, native 22.05 kHz *retained*, silences over 2 s trimmed at 40 dB, duration normalised to 4 s by centre-cropping or reflection padding. Retaining the native rate is a deliberate departure from our prior pipeline: downsampling to 16 kHz imposes an 8 kHz Nyquist ceiling and removes the 8–11 kHz band, deleting the evidence most likely to distinguish a GAN vocoder from an end-to-end model. All filterbanks run to $f_{\max} = 11.025$ kHz, and each feature map is standardised per

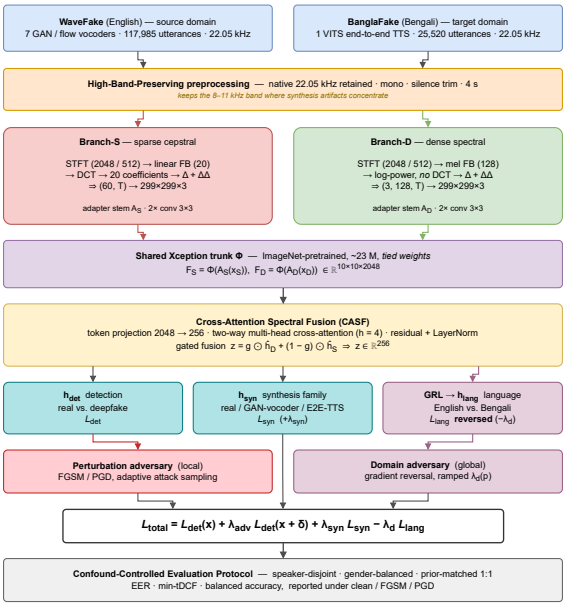

Fig. 1. System architecture and end-to-end pipeline of DASF-Net. Both corpora pass through identical High-Band-Preserving preprocessing, split into a sparse cepstral and a dense spectral branch, and are encoded by one Siamese-shared Xception trunk. CASF produces a single embedding z on which three heads act, and two adversaries shape that same embedding.

utterance and squashed to $[0, 1]$, so ϵ means the same thing everywhere.

Two pipelines share identical STFT parameters ($n_{\text{fft}} = 2048$, hop $= 512$), so any difference is attributable to the filterbank alone. **Branch-S** (sparse cepstral): linear filterbank (20 filters, 0–11.025 kHz) $\rightarrow$ DCT $\rightarrow$ 20 cepstral coefficients $\rightarrow \Delta + \Delta\Delta$, giving $(60, T)$. **Branch-D** (dense spectral): mel filterbank ($n_{\text{mels}} = 128$) $\rightarrow$ log-power, with *no* DCT so the dense time–frequency structure survives $\rightarrow \Delta + \Delta\Delta$, giving $(3, 128, T)$. Both are resized to $299 \times 299 \times 3$.

C. System Architecture

Adapters and a Siamese-shared trunk. Each branch owns a small *adapter stem* A_S, A_D — two 3×3 convolutions with batch normalisation and ReLU, about 0.06 M parameters each — mapping branch-specific statistics into a common range. The encoder Φ is a single ImageNet-pretrained Xception [10] ($\approx$ 23 M) applied to both branches with tied weights:

$$F_S = \Phi(A_S(x_S)), \qquad F_D = \Phi(A_D(x_D)), \tag{1}$$

with $F_S, F_D \in \mathbb{R}^{10\times10\times2048}$. Sharing is a design commitment, not an economy: two independent backbones would each build a private feature space, leaving the fusion module to reconcile two unrelated geometries. One trunk forces both views into a single artifact space for 0.12 M extra parameters.

Cross-Attention Spectral Fusion. Each feature map is flattened to 100 tokens and projected to $d = 256$, giving $T_S, T_D \in \mathbb{R}^{100\times256}$. Two multi-head cross-attention blocks [12] ($h = 4$) let each view interrogate the other,

$$\hat{h}_D = \text{LN}(T_D + \text{MHA}(T_D, T_S, T_S)), \tag{2}$$

$$\hat{h}_S = \text{LN}(T_S + \text{MHA}(T_S, T_D, T_D)), \tag{3}$$

and a learned gate mixes them before mean-pooling over tokens:

$$g = \sigma\Big(W_g[\hat{h}_D; \hat{h}_S] + b_g\Big), \tag{4}$$

$$z = \text{pool}\Big(g \odot \hat{h}_D + (1 - g) \odot \hat{h}_S\Big) \in \mathbb{R}^{256}. \tag{5}$$

The gate is the point of the module. Concatenation would assert both views always matter equally; the gate lets the network decide per utterance and per dimension which representation carries the artifact, and makes the model interpretable — the distribution of g answers directly whether GAN vocoders and end-to-end TTS leave traces in different representations.

Objective heads. Three heads read z. h_{det} ($256 \rightarrow 128 \rightarrow 2$) performs the task of record under cross-entropy with label smoothing 0.1. h_{syn} ($\rightarrow 3$) predicts {real, GAN-vocoder, end-to-end TTS}; this is deliberately coarse, because predicting the specific generator would invite memorisation of generator identity, exactly the failure that breaks cross-vocoder transfer. h_{lang} predicts English versus Bengali but sits behind a gradient reversal layer (GRL) [11]: identity forward, gradient multiplied by $-\lambda_d$ backward. The head does its best to identify the language while the trunk and CASF are driven to make that impossible.

Dual-adversarial objective. With δ an adversarial perturbation,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}}(x) + \lambda_{\text{adv}}\mathcal{L}_{\text{det}}(x+\delta) + \lambda_{\text{syn}}\mathcal{L}_{\text{syn}} - \lambda_d\mathcal{L}_{\text{lang}}, \tag{6}$$

the negative sign realised by the GRL rather than by ascending the loss. The perturbation adversary is *local*, searching a small neighbourhood of each input for a direction that flips the decision; the domain adversary is *global*, searching the embedding for any direction revealing the corpus. We hypothesise they are

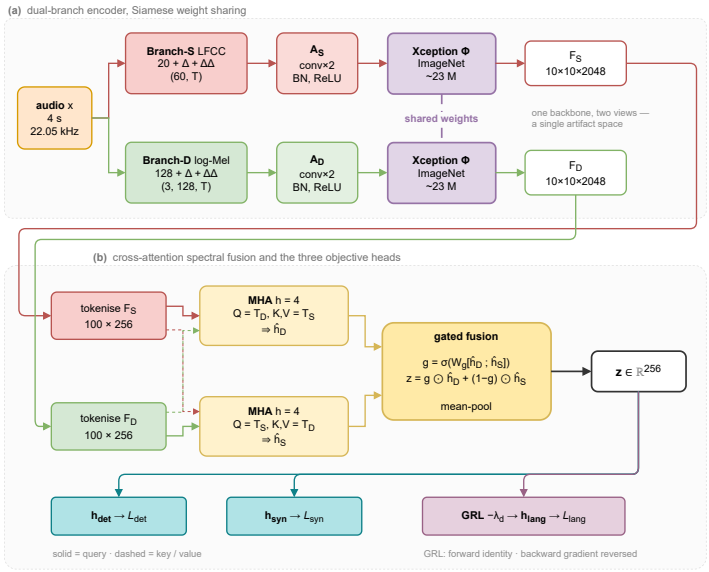

Fig. 2. DASF-Net in detail. (a) The dual-branch encoder, each view passed through its own adapter stem into a single weight-shared Xception trunk. (b) CASF tokenises both feature maps, lets each view attend over the other, and combines them through a learned gate into z; the language head sits behind a gradient reversal layer.

complementary, and the ablation below tests that rather than assuming it.

D. Training Pipeline

Training (6) from scratch is unstable: a gradient-reversed discriminator on an untrained encoder produces a large uninformative reversed gradient, and adversarial examples against an untrained boundary are meaningless. Difficulty is therefore staged (Fig. 3). **Stage 1 — warm-up:** Φ frozen at ImageNet weights; only A_S, A_D, CASF and h_{det} train on labelled source data under $\mathcal{L}_{\text{det}}$; Adam, $lr = 10^{-4}$, effective batch 128. **Stage 2 — alignment:** Φ unfreezes and h_{syn}, h_{lang} attach; batches mix labelled source with *unlabelled* target audio, target labels never used, only the language tag, keeping the setting honestly unsupervised on the target. Following [11] the reversal coefficient is ramped, $\lambda_d(p) = \lambda_{\max}(2/(1+e^{-10p}) - 1)$, p being the fraction of the stage completed. **Stage 3 — hardening:** all modules fine-tune at $lr = 5 \times 10^{-5}$ for 10 epochs with the language adversary still active; the attack per minibatch is sampled in proportion to a momentum-smoothed success estimate, $s_a \leftarrow (1-m)s_a + m\hat{s}_a$ with $m = 0.2$, while $p_c = 1/3$ of each minibatch stays clean to preserve baseline accuracy.

E. Threat Model, Tools and Resources

Two white-box attacks are evaluated in the normalised feature domain. FGSM [13] takes one step $\delta = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}_{\text{det}})$; PGD [14] iterates $K = 10$ steps at $\alpha = \epsilon/4$, projecting onto the ℓ_∞ ball each step. Critically, *both are swept over the same grid*, $\epsilon \in \{0.001, 0.005, 0.010\}$: prior work, including our own, reports FGSM and PGD at ϵ values two orders of magnitude apart, which makes those robustness figures non-comparable.

Implementation uses Python 3 and PyTorch, `timm` for the ImageNet-pretrained Xception, `librosa`/`soundfile`/`scipy` for audio, `scikit-learn` for ROC/EER and `numpy`/`pandas` for data handling; attacks are written directly in PyTorch autograd, diagrams in draw.io and the report in LaTeX. All software is open source and both datasets public. Training runs on one local NVIDIA RTX 5060 Ti with 8 GB VRAM, a binding constraint: the shared trunk is evaluated twice per utterance, so one sample costs roughly double the activation memory of a single-branch Xception at 299×299 and a batch of 128 does not fit. Training therefore uses a **micro-batch of 16 with gradient accumulation over 8 steps**, giving an effective batch of 128 and preserving the optimisation behaviour the learning rates assume, at the cost of wall-clock time. Mixed precision (FP16) is used throughout.

V. Testing Methodology

Metrics. The headline metric is **Equal Error Rate (EER)**, the ROC point where false positive and false negative rates are equal. EER is threshold-free, which matters because the source corpus has a roughly 8:1 synthetic-to-real prior and the target roughly 1:1, so a threshold calibrated on one is miscalibrated on the other. **Balanced accuracy** accompanies it; raw accuracy is never reported for a transfer condition without its prior. min-tDCF is *not* reported: it is defined over a tandem countermeasure and ASV system, and neither corpus ships ASV scores.

Benchmarks. ResNet18 fine-tuned on BanglaFake at 0.2435 EER [8]; single-branch Xception with log-Mel at 0.034 average EER on WaveFake [9]; GMM+LFCC at 0.058 EER [6].

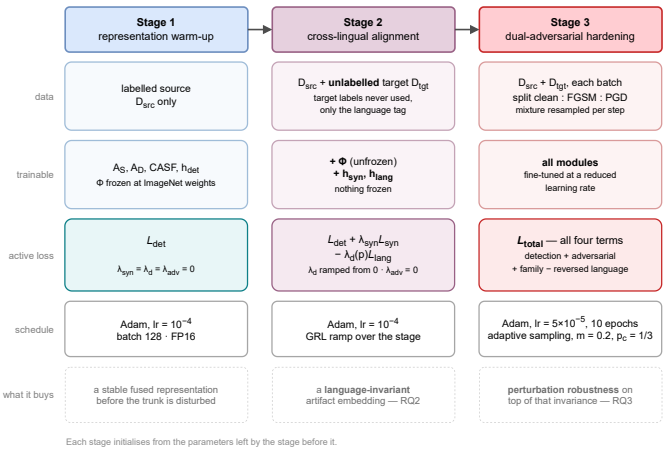

Fig. 3. The three-stage training curriculum. Each stage adds one source of difficulty and inherits the weights of the stage before it.

TABLE II
PLANNED EXPERIMENTS AND THE OBJECTIVE EACH ADDRESSES

ID	Experiment	Obj.
E1	In-language baselines: Branch-S only, Branch-D only, full CASF fusion	2
E2	Zero-shot transfer both directions, with and without the language adversary; per-vocoder breakdown; $\leq$10% few-shot	3
E3	Clean, FGSM and PGD EER per ϵ; in-domain vs. transferred	4
E4	**Ablation**: remove CASF, shared trunk, h_{syn}, GRL, adversarial stage, or HBP, one at a time	1–4
E5	Gate analysis: distribution of g per synthesis family	1

TABLE III
PROJECT TIMELINE AND MILESTONES

Ph.	Activity	Status
1	Literature review; dataset acquisition and manifests (May–Jun 2026)	Complete
2	Architecture design; algorithms; diagrams; implementation (Jun–Aug 2026)	Complete
3	Report preparation and submission (Aug–6 Sep 2026)	Complete
4	Environment setup; training runs E1–E3 (from 6 Sep 2026)	In progress
5	Ablation E4, gate analysis E5; results write-up	Pending

Confound-Controlled Evaluation Protocol. Every reported figure obeys four constraints, following the confounds of Section III: (1) *speaker-disjoint splits*, so no model can succeed by speaker recognition; (2) *gender-balanced Bengali evaluation* — because BanglaFake pairs one male synthetic speaker against mixed-gender real speech, the primary Bengali test set is restricted to male real speech and the mixed-gender result reported separately as a diagnostic, a large gap between them being itself evidence of gender shortcutting; (3) *prior-matched test sets*, both resampled to 1:1; and (4) EER as the threshold-free primary metric.

Experiments. Five are planned (Table II). E4 is what converts an observed improvement into an attributed one. All runs use three random seeds reported as mean and standard deviation. Objective 1 is met when the implementation passes its correctness tests (tensor shapes, GRL sign flip, attack budget compliance); Objectives 2–4 when E1–E3 produce seed-averaged EER under CCEP, whatever those figures turn out to be. A negative result — for instance that the language adversary does not improve transfer — satisfies the objective provided the protocol could have detected an improvement.

Monitoring. Progress is tracked through weekly supervisor meetings, a shared Git repository holding all code and experiment configurations, and per-epoch logs recording each loss term separately plus running attack-success estimates. Separate loss logging is what makes the GRL diagnosable: a language loss collapsing toward chance means the adversary works, one collapsing toward zero means it has overwhelmed the encoder.

Current status. No experiments have been executed. The architecture, training algorithms, threat model and evaluation protocol are complete and the reference implementation is written, but it has not been trained; this report therefore contains no results and no trained model. Reporting this plainly is preferred to presenting an untested design as a validated one.

VI. Project Timeline and Budget

Table III lists the phases. Phases 4 and 5 fall after the report deadline of 6 September 2026: design and implementation were completed within the reporting period, the training runs were not. The critical path is Phase 4 — the dependency install and first full training run must both succeed before Phase 5 can begin.

Training runs on a locally owned RTX 5060 Ti, so direct monetary cost is negligible; the real cost is time and electricity (Table IV). Demand is driven by the trunk running twice per utterance and Stage 3 PGD at $K = 10$, and the estimate assumes on the order of 200 GPU-hours over all experiments and three seeds, with order-of-magnitude figures assuming 1 USD $\approx$ 120 BDT, 8 BDT/kWh and about 0.2 kW draw. The binding resource is the 8 GB VRAM ceiling and the wall-clock time micro-batching implies; the contingency line is not planned spend but the cost incurred only if the local card cannot finish in time.

TABLE IV
ESTIMATED PROJECT BUDGET

Item	USD	BDT
Personnel (2 students, academic credit)	0	0
GPU hardware (RTX 5060 Ti, already owned)	0	0
Electricity, ≈200 GPU-h @ 0.2 kW	3	320
Storage, software, datasets (local / open / public)	0	0
Total, direct cost	**3**	**320**
Contingency: cloud fallback (≈200 h)	*300*	*36,000*

TABLE V
RISKS AND MITIGATION STRATEGIES

Risk	Mitigation
R1 GRL divergence	Enters only in Stage 2, never at initialisation, λ_d ramped from zero; language loss logged separately so divergence is visible. Fallback: maximum mean discrepancy alignment, reported as a negative result rather than concealed
R2 Gender confound	CCEP restricts the primary Bengali test set to male real speech and reports the mixed-gender figure separately; a large gap is read as shortcutting, not success
R3 Memory ceiling	Micro-batch 16 with accumulation over 8 steps holds the effective batch at 128; WaveFake subsampled to ≈26,000 utterances — on this hardware a necessity, not an option; PGD at $K = 5$ for ablations, full budget kept for E3; gradient checkpointing in reserve
R4 One generator	Claims scoped explicitly to cross-*paradigm* transfer and the limitation stated openly; adding a Bengali FastSpeech2 or Tacotron2 is the next step
R5 Schedule slip	E1 and E2, which answer Objectives 2 and 3, run first; E4 and E5 are deferrable. A smoke test on random tensors validates the code before compute is committed
R6 Dependencies	The implementation accepts an injected backbone, so it can be validated with a lightweight stand-in trunk before the full stack is installed

VII. TEAM MEMBERS AND ROLES

The two student members contributed equally, taking ten report sections each. **Iftikhar Ahmed** (2121019642) — model and method: architecture design (CASF, the dual-adversarial objective, the curriculum), implementation, threat model, testing methodology, risk analysis, diagrams and LaTeX preparation. **Meherun Nessa Mithila** (2221649042) — data and evaluation: literature and paper reviews, dataset acquisition and manifests, preprocessing pipeline, evaluation protocol, timeline and budget, and language editing. **Mohammad Abdul Qayum**, Assistant Professor, ECE — faculty and supervisor: scope definition and methodological review.

VIII. RISKS AND MITIGATION STRATEGIES

Table V lists the six principal risks. R3 is the most probable cause of the project failing to produce results, and R1 the most probable cause of the method itself not working.

IX. CONCLUSION AND LIMITATIONS

This project proposes DASF-Net, an architecture treating the two standing weaknesses of audio deepfake detectors — brittleness across languages and brittleness under perturbation — as instances of one problem, attacked with adversaries acting on a single shared embedding. It contributes a dual-branch spectral front-end over a Siamese-shared Xception trunk, a gated cross-attention fusion module, a combined objective in which a gradient-reversed language discriminator and adaptive FGSM/PGD training shape the same representation, a three-stage curriculum making this trainable, a corrected threat model, and an evaluation protocol neutralising the confounds that otherwise make Bengali detection results unreadable. It addresses two problems the most recent Bengali benchmark [8] named as open, with a mechanism rather than more data or a larger model. If the hypothesis holds, the result is a detector whose features describe what a synthesiser did rather than which corpus it was trained on — a property generalising to any low-resource language with unlabelled speech.

Limitations, stated in advance rather than discovered later. *No results yet:* the experiments of Section V have not been executed, so every claim here is a design claim. *One Bengali generator:* the design tests cross-paradigm but not within-Bengali cross-vocoder generalisation, and no such claim is made. *Confounds mitigated, not eliminated:* each CCEP control shrinks the usable evaluation set, trading power for validity. *White-box attacks only:* black-box, transfer, perceptual and physical-channel attacks are outside this design. *Two languages:* Bengali and English differ in phoneme inventory, prosody and recording conditions at once, so invariance over one pair is evidence, not proof. *Computational cost:* Stage 3 roughly doubles training time on top of the two-branch encoder's already doubled forward cost.

REFERENCES

[1] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.
[2] K. Kumar *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS*, 2019.
[3] R. J. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, 2019.
[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021.
[5] M. Todisco *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, 2019.
[6] J. Frank and L. Schönherr, "WaveFake: A data set to facilitate audio deepfake detection," in *NeurIPS Datasets and Benchmarks Track*, 2021.
[7] I. A. Fahad, K. Asif, and S. Sikder, "BanglaFake: Constructing and evaluating a specialized Bengali deepfake audio dataset," *arXiv:2505.10885*, 2025.
[8] M. S. S. Samu *et al.*, "Zero-shot to zero-lies: Detecting Bengali deepfake audio through transfer learning," in *Proc. ICCIT*, 2025.
[9] I. Ahmed, M. I. Arian, and N. J. Lisa, "Mel-Spectrogram representations for CNN based adversarial defense for audio deepfake detection," in *Proc. ICMLT*, Berlin, Germany, 2026.
[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE CVPR*, 2017.
[11] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, 2016.
[12] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS*, 2017.
[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.
[14] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018.
[15] P. Kawa, M. Plata, and P. Syga, "Defense against adversarial attacks on audio deepfake detection," in *Proc. Interspeech*, 2023.
[16] K. Ito and L. Johnson, "The LJ speech dataset," 2017.
[17] A. Ahmad, M. Selim, M. Iqbal, and M. Rahman, "SUST TTS corpus: A phonetically balanced corpus for Bangla text-to-speech synthesis," *Acoust. Sci. Technol.*, vol. 42, 2021.
[18] R. Ardila *et al.*, "Common Voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020.