A Unified Feature Mixture Framework for Joint Speech and Singing Deepfake Detection

Aastha Sharma and Guangjing Wang
University of South Florida, Tampa, FL, USA

Abstract

High-fidelity audio generation techniques, such as voice conversion and singing voice synthesis, have significantly increased the risk of audio deepfakes. Although existing methods perform well on conversational speech deepfake detection, they fail severely under the speech-to-singing domain shift. To address this limitation, we propose GenuVoice, a unified deepfake detector based on a multi-branch mixture-of-experts architecture that integrates three complementary feature views: Wav2Vec 2.0 representations, log-mel spectrograms, and mel-frequency cepstral coefficients (MFCC). Each expert is trained to remain independently discriminative, while a learned gating network dynamically weights expert contributions. A speech-retentive multi-domain fine-tuning strategy enables adaptation to singing without degrading speech performance. GenuVoice achieves 1.82% Equal Error Rate (EER) on CtrSVDD, compared to 37–62% for existing speech-trained detectors, while preserving strong speech performance (0.38% EER on ASVspoof 2019) and generalizing to unseen generators (8.89% EER on held-out ASVspoof 2021). Extensive ablations confirm the importance of multi-expert fusion and speech retention, establishing GenuVoice as an effective unified detector for speech and singing deepfakes. The implementation code is available at https://github.com/aastha-sharma/genuvoice

1 Introduction

The rapid advancement of generative models has enabled high-fidelity text-to-speech (TTS), voice conversion (VC), and singing voice synthesis (SVS) (Liu et al., 2022; Zhang et al., 2022). While these technologies offer substantial practical benefits, they also introduce significant security and ethical risks, including identity impersonation and intellectual property infringement.

To date, audio forensics research has primarily focused on conversational speech. In contrast, the growing prevalence of AI-generated music and unauthorized artist voice cloning expands the threat landscape beyond speech, exposing a critical gap in existing detection systems. Existing audio deepfake detection methods often fail to generalize reliably across both speech and singing domains (Zang et al., 2024b,a; Gohari et al., 2025). Modern media streams often contain both speech and singing within the same content (e.g., spoken intros around sung segments). Treating speech and singing as entirely separate tasks requires maintaining two detectors and an additional domain routing component, which increases deployment complexity and introduces failure modes when routing is incorrect. This gap underscores the need for a unified detector that operates reliably across both speech and singing using a single inference pipeline.

Existing speech-specific detectors (Tak et al., 2021; Jung et al., 2022) achieve strong performance on benchmark datasets such as ASVspoof 2019 (Todisco et al., 2019) or ASVspoof 2021 (Liu et al., 2023). In contrast, singing-specific deepfake detection remains underexplored and is constrained by limited labeled data and heterogeneous recording conditions (Zang et al., 2024b,a; Gohari et al., 2025). The fundamental challenge is the acoustic domain gap: singing exhibits sustained harmonics, vibrato, extended pitch ranges, and complex timbre variations that are absent in conversational speech (Gohari et al., 2025). Sequential adaptation from speech to singing can therefore induce catastrophic forgetting (Kirkpatrick et al., 2017), where fine-tuning overwrites robust speech representations while providing limited gains on singing.

To fill the gap, we propose GenuVoice, a simple but effective framework with adaptive mixtures of local experts (Jacobs et al., 1991) to achieve unified speech and singing deepfake detection. Specifically, we choose three complementary feature types using Wav2Vec 2.0, log-mel spectrogram, and mel-frequency cepstral coefficients (MFCC)

Findings of the Association for Computational Linguistics: ACL 2026, pages 24853–24863

based on an existing benchmark study for audio features investigation (Gohari et al., 2025). The self-supervised speech representations from Wav2Vec 2.0 capture rich contextual structure information for speech spoofing (Baevski et al., 2020; Tak et al., 2022), the log-mel spectrogram features are particularly effective for singing artifacts and harmonic irregularities (Gohari et al., 2025), and MFCC remains a strong and computationally efficient baseline for anti-spoofing (Todisco et al., 2019; Liu et al., 2023).

GenuVoice is built upon a three-stage pipeline: (i) training an individual expert model (e.g., ResNet-18) only on speech features; each branch is optimized with an auxiliary classification loss; (ii) training a gating network model and final classifier (e.g., Multilayer Perceptron) while freezing the expert models on speech features. The gating network adaptively assigns feature weights to the classifier input, and the classifier outputs the final deepfake detection result. The first two stages can be regarded as an initialization stage to avoid cold-start issues; and (iii) multi-domain fine-tuning the expert models, the gating network model, and the final classifier on a mixture of speech and singing features. This mechanism maintains speech-related representations while facilitating the acquisition of singing-specific spoofing cues.

We evaluate GenuVoice through a series of experiments spanning domain generalization, expert contribution analysis, baseline comparisons, and robustness to domain alignment strategies. We show that existing speech-trained detectors (e.g., RawNet2, AASIST, Wav2Vec2-AASIST, Wav2Vec2-DF) all achieve 37–62% EER on the CtrSVDD singing benchmark, confirming catastrophic cross-domain failure. GenuVoice achieves 1.82% EER on CtrSVDD, a 95% relative improvement over the best baseline, while preserving strong speech performance (0.38% EER on ASVspoof 2019) and generalizing to unseen generators and codec conditions (8.89% EER on held-out ASVspoof 2021). We further show that adversarial domain alignment (DANN) induces negative transfer even under matched training conditions (26.45% EER at 50/50 mixing), validating our supervised multi-domain approach. Ablation studies confirm that removing auxiliary expert supervision or diversity regularization substantially degrades performance, increasing EER to 3.45–12.34%. In summary, we make the following contributions:

- We propose a mixture-of-experts architecture for unified speech and singing deepfake detection, where complementary feature experts are jointly modeled with auxiliary supervision to remain independently discriminative, and a learned gating network enables adaptive specialization across vocal modalities.
- We introduce a speech-retentive multi-domain fine-tuning pipeline that jointly optimizes speech and singing data to mitigate catastrophic forgetting, allowing the model to acquire singing-specific spoofing cues while preserving strong speech-domain performance within a single inference framework.
- We conduct a systematic evaluation spanning zero-shot transfer, multi-domain fine-tuning, expert ablations, and adversarial domain alignment, and provide direct comparisons against four published speech-trained detectors and the most relevant prior singing detection work, quantifying the severity of speech-to-singing generalization failure and confirming that GenuVoice substantially outperforms all baselines on singing while retaining speech-level performance.

2 Related Work

2.1 Speech Deepfake Detection

Progress in audio deepfake detection has historically been driven by curated benchmarks such as ASVspoof (Todisco et al., 2019; Liu et al., 2023), which standardize attack types, splits, and evaluation protocols (Wang et al., 2023b; Guo et al., 2023; Wang et al., 2025). The ASVspoof 2019 Logical Access (LA) benchmark remains a widely used testbed for evaluating detection under TTS and voice conversion attacks. While these benchmarks enabled rapid improvement under in-domain settings, existing studies show that performance can degrade substantially under distribution shift and real-world post-processing (Müller et al., 2022).

Existing work explores features and inductive biases designed to generalize beyond purely data-driven deepfake classifiers. For example, Kumari et al. (2025) proposed VoiceRadar, which incorporates physically inspired modeling of subtle frequency dynamics (micro-frequencies) to provide complementary cues for deepfake detection. This discovery argues that certain fine-grained dynamics present in human-produced audio may be dif-

ficult for generative models to reproduce consistently. Our multi-expert approach aligns with the broader motivation of combining complementary cues. Rather than introducing a new handcrafted feature family, we fuse heterogeneous learned and hand-engineered representations (self-supervised, spectral, and cepstral) within a single trainable framework for audio deepfake detection.

Recent work explored mixture-of-experts (MoE) architectures for speech deepfake detection. Negroni et al. (2025) leverage MoE to fuse complementary features for improved speech-only detection, while Hao et al. (2025) propose a hierarchical expert fusion with two-stage learning for robust speech deepfake detection. Both approaches apply MoE within the speech domain to improve in-domain performance. In contrast, our work addresses a different challenge: cross-domain generalization between speech and singing, where spoof cues are structurally different across domains. Our multi-view MoE formulation, combined with a speech-retentive adaptation pipeline, is specifically designed to handle the acoustic heterogeneity between vocal domains.

2.2 Singing Deepfake Detection

Compared to speech, singing voice deepfake detection remains underexplored despite the increasing availability of singing synthesis and singing voice conversion systems. For instance, Zang et al. (2024b) curated the SingFake dataset from user-generated platforms and showed that strong speech countermeasures can fail catastrophically on singing mixtures. Gohari et al. (2025) performed a systematic feature study for singing deepfake detection and reported that time-frequency representations such as log-mel spectrograms can be more robust than waveform encoders on singing audio, likely because they capture harmonic inconsistencies more directly in the presence of accompaniment. To facilitate controlled evaluation, the SVDD benchmark was introduced to study singing deepfake detection under more standardized conditions (Zang et al., 2024a). However, most prior work treats speech and singing as separate detection tasks or evaluates speech detectors on singing in a zero-shot manner. In contrast, our work investigates *speech-retentive multi-domain fine-tuning* of a speech-trained detector for unified speech and singing detection, explicitly measuring the trade-off between preserving strong speech performance while improving singing detection.

3 System Design

In this section, we introduce the design assumption and our designed model for unified speech deepfake and singing deepfake detection.

3.1 Assumption

Given an input audio waveform $x \in \mathbb{R}^T$ sampled at 16 kHz, the goal is to learn a function $f(x) \rightarrow [0, 1]$ that predicts the probability of the audio being a deepfake. During the inference stage, predictions should be accurate for samples drawn from either the speech or singing distributions.

The core assumption for our designed framework is that no single feature representation is optimal for both the speech domain and the singing domain. Unlike typical domain adaptation scenarios where source and target distributions differ primarily in marginal statistics, speech and singing exhibit fundamentally different structural properties. Speech deepfakes often manifest artifacts tied to temporal dynamics and phonetic structure, whereas singing deepfakes are characterized by harmonic distortions, vibrato inconsistencies, and unnatural pitch dynamics (Gohari et al., 2025; Zang et al., 2024b). This divergence weakens the shared-representation assumption underlying adversarial alignment.

In addition, fine-tuning a speech-trained detector on limited singing data can destabilize previously learned speech representations (French, 1999; Kirkpatrick et al., 2017). The cross-domain fine-tuning will produce a tension between preserving speech deepfake detection robustness and improving singing deepfake detection. This is especially problematic because speech deepfake features are different from the singing deepfakes as discussed above. Besides, publicly available singing datasets are typically smaller and more heterogeneous than speech benchmarks (Zang et al., 2024a,b), providing insufficient supervision to recover lost source-domain competence after aggressive adaptation.

3.2 Expert Feature Selection

To tackle the acoustic heterogeneity between speech and singing, we propose a Multi-Expert Unified Architecture as shown in Figure 1. We combine three complementary views: spectral (log-mel), cepstral (MFCC), and self-supervised (Wav2Vec 2.0), and learn to fuse them adaptively with a gating network.

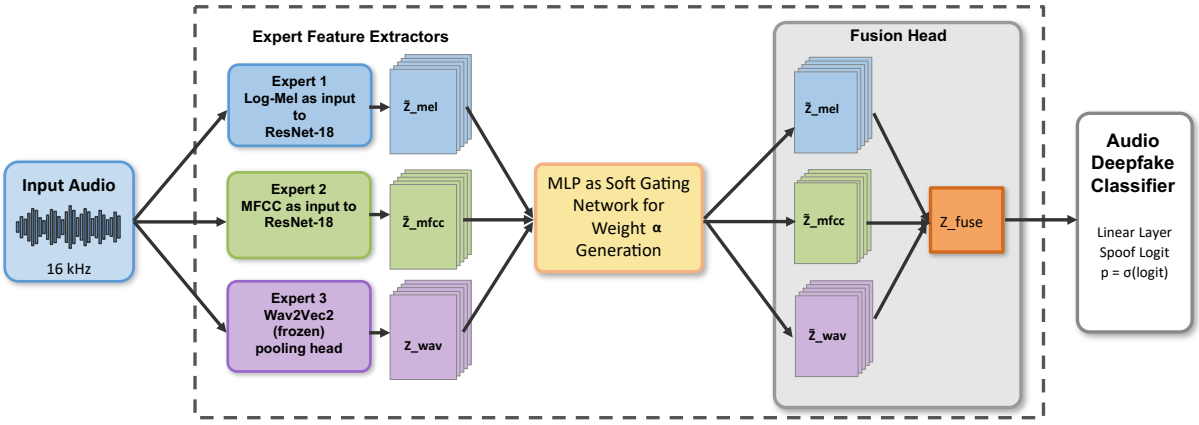

Figure 1: GenuVoice audio deepfake detector architecture. Expert modules produce heterogeneous meta-embeddings. The fusion head projects each embedding to a shared dimension with temperature-scaled mixture weights.

3.2.1 Log-Mel Spectrogram

Log-mel spectrograms provide a dense time-frequency representation that highlights harmonic structure. Prior work (Gohari et al., 2025; Zang et al., 2024b) finds that spectrogram features can be more robust than waveform encoders on singing audio, likely because they expose harmonic inconsistencies that persist even with accompaniment. We extract Log-Mel spectrograms with 128 Mel bins using a 25 ms window and 10 ms hop at 16 kHz, followed by log compression. The resulting single-channel time–frequency representation is processed by a *ResNet-18* encoder, trained with a cross-entropy objective, and projected into a 512-dimensional meta-embedding $\mathbf{z}_{\text{mel}} \in \mathbb{R}^{512}$.

3.2.2 MFCC

MFCCs compress the spectral envelope into a low-dimensional cepstral space and remain strong, efficient baselines for anti-spoofing in the MFCC/LFCC family (Todisco et al., 2019; Liu et al., 2023). MFCCs can capture envelope irregularities and discontinuities introduced by vocoders and conversion systems. We extract 40 MFCC coefficients using the same 25 ms / 10 ms windowing as the log-mel branch. MFCC features are processed by an independent *ResNet-18* encoder, trained with a cross-entropy objective, producing a 512-dimensional meta-embedding $\mathbf{z}_{\text{mfcc}} \in \mathbb{R}^{512}$.

3.2.3 Wav2Vec 2.0

The Wav2Vec 2.0 learns contextualized representations via self-supervised pretraining (Baevski et al., 2020) and is effective for deepfake detection in speech (Tak et al., 2022). Wav2Vec 2.0 audio representation captures higher-level temporal irregularities that handcrafted features may miss. We use the `wav2vec2-base-960h` checkpoint. Given a 4-second crop at 16 kHz, the model produces frame-level hidden states of shape $T \times 768$, where T denotes the number of latent frames produced by the Wav2Vec2 feature encoder for the crop. We reduce dimensionality using a lightweight 1D-CNN head with two convolutional blocks ($768 \rightarrow 256 \rightarrow 128$), followed by global average pooling over T latent frames, yielding $\mathbf{z}_{w2v} \in \mathbb{R}^{128}$.

3.3 Adaptive Gating and Fusion

The gating network is a multilayer perceptron (MLP) model, which combines different feature embeddings using soft mixture weights. The input of MLP is the concatenated feature embeddings, and the output is the weights. We concatenate expert embeddings: $\mathbf{u} = [\mathbf{z}_{mel} \oplus \mathbf{z}_{mfcc} \oplus \mathbf{z}_{w2v}] \in \mathbb{R}^{1152}$, and feed $\mathbf{u}$ into an MLP to obtain unnormalized gate logits $\mathbf{g} \in \mathbb{R}^3$. Mixture weights are computed using a temperature-scaled softmax: $\boldsymbol{\alpha} = \text{softmax}(\mathbf{g}/\tau)$, where τ controls the sharpness of expert weighting and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3]$.

As expert features have different dimensionalities, we apply learnable linear projections to map each expert to a common fusion space $d = 128$: $\tilde{\mathbf{z}}_i = W_i \mathbf{z}_i, W_i \in \mathbb{R}^{128 \times d_i}$, where $d_i \in \{512, 512, 128\}$ is the original expert dimension. We then compute the fused embedding as a weighted sum:

$$\mathbf{z}_{\text{fuse}} = \sum_{i=1}^{3} \alpha_i \tilde{\mathbf{z}}_i \in \mathbb{R}^{128}. \quad (1)$$

The $\mathbf{z}_{\text{fuse}}$ is used as input to a final linear classification head for audio deepfake detection.

3.4 Optimization Objectives

A common failure mode in adaptive mixtures of local experts is *gate collapse*, where the gate assigns almost all weight to a single expert (Jacobs et al., 1991; Jordan and Jacobs, 1994). We mitigate this using (i) auxiliary supervision for each expert, (ii) entropy regularization on gate weights, and (iii) a diversity penalty that encourages distinct expert representations. Specifically, let $\hat{y}$ be the fused prediction and l_i the prediction from the i-th expert's auxiliary classifier. The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} - \lambda_{\text{ent}}\mathcal{H}(\boldsymbol{\alpha}) + \lambda_{\text{div}}\mathcal{L}_{\text{div}}, \quad (2)$$

$$\mathcal{H}(\boldsymbol{\alpha}) = -\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{3}\alpha_{b,i}\log(\alpha_{b,i}+\epsilon), \quad (3)$$

where $\mathcal{L}_{\text{task}} = \text{CE}(y, \hat{y})$ is the primary cross-entropy loss on the fused output. $\mathcal{L}_{\text{aux}} = \sum_{i=1}^{3} w_i \cdot \text{CE}(y, l_i)$ encourages each expert to remain independently discriminative. We empirically set $\lambda_{\text{aux}} = 0.1$. $\mathcal{H}(\alpha)$ is the average gate entropy over a minibatch of size B. The term $-\lambda_{\text{ent}}\mathcal{H}$ encourages higher entropy (more diverse expert usage). We set $\lambda_{\text{ent}} = 0$ initially to allow the gate to identify strong experts, then enable $\lambda_{\text{ent}} = 10^{-4}$ at later stages. $\mathcal{L}_{\text{div}} = \frac{1}{3}\sum_{i<j}\text{cos-sim}(\mathbf{z}_i, \mathbf{z}_j)$ penalizes cosine similarity between expert features, encouraging diverse representations. We set $\lambda_{\text{div}} = 0.1$ throughout training.

4 Experimental Setup

4.1 Datasets

Table 1 summarizes the dataset partitions used in our pipeline. We group datasets by domain: *speech* (ASVspoof 2019, WaveFake, ASVspoof 2021) and *singing* (CtrSVDD, SingFake). Dataset usage across stages is as follows:

- **Stage 1–2 (speech-only training):** ASVspoof 2019 (train/dev) and WaveFake (train/dev) are merged into a single speech training pool. Validation and early stopping are performed on the merged **speech dev** pool.
- **Stage 3 (multi-domain fine-tuning):** we fine-tune using a balanced 50/50 mixture of speech (ASVspoof 2019 train + WaveFake train) and singing (CtrSVDD train + SingFake train), enforced at the batch level.
- **Testing (held-out evaluation):** we report results on the official test partitions of ASVspoof 2019, WaveFake, CtrSVDD, and SingFake. In addition, we evaluate on ASVspoof 2021, which is held out entirely from training, to assess generalization to unseen generators and codec conditions.

Table 1: Dataset partitions. ASVspoof 2021 is used exclusively for held-out evaluation and is not included in any training stage. Sources: ASVspoof 2019 (Todisco et al., 2019), ASVspoof 2021 (Liu et al., 2023), WaveFake (Frank and Schönherr, 2021), CtrSVDD (Zang et al., 2024a), SingFake (Zang et al., 2024b).

Dataset	Domain	Train	Dev	Test
ASVspoof 2019	Speech	25,380	24,844	18,029
WaveFake	Speech	8,734	1,093	4,390
CtrSVDD	Singing	8,236	203	12,326
SingFake	Singing	563	18	138

4.2 Audio Preprocessing

All audio is converted to 16 kHz mono PCM and peak-normalized to $[-1, 1]$. For singing corpora (CtrSVDD, SingFake), we perform vocal separation using Demucs (Défossez et al., 2019). We log vocal separation outcomes per file and report the number of fallback samples for reproducibility.

All models operate on fixed-length 4-second segments (64,000 samples at 16 kHz). During training, files longer than 4 seconds are randomly cropped; files shorter than 4 seconds are zero-padded at the end. During evaluation, we use multi-crop averaging: for each file longer than 4 seconds, we extract $K = 5$ evenly spaced crops and average the predicted spoof posterior across crops. Crop offsets are deterministic under a fixed seed.

During multi-domain fine-tuning, each batch is constructed to be domain-balanced: 50% speech samples and 50% singing samples (CtrSVDD + SingFake). When a pool is smaller (notably SingFake), we sample with replacement to maintain a strict 50/50 ratio per batch. This prevents the optimizer from being dominated by the larger speech corpora and stabilizes gate recalibration.

4.3 Training Stages

Stage 1: Backbone warmup (speech only). We train the expert models with auxiliary classification heads. We boost the Wav2Vec auxiliary weight to 1.5 (vs. 1.0 for Log-Mel/MFCC) to compensate for the slower convergence of self-supervised representations. Temperature uses the default $\tau = 1.0$ (no scaling), and entropy regularization is disabled.

Table 2: Zero-shot evaluation. GenuVoice trained only on the speech dataset fails to generalize to singing.

Dataset	Domain	Samples	EER (%)
In-Domain (Speech)			
ASVspoof 2019	Speech	18,029	**0.38**
WaveFake	Speech	4,390	0.05
Out-of-Domain (Singing)			
SingFake	Singing	138	53.63
CtrSVDD	Singing	12,326	43.16
Pooled Singing	**Singing**	**12,464**	**43.25**

Stage 2: Partial unfreeze (speech only). We freeze the ResNet backbones and use the lightweight 1D-CNN pooling head on Wav2Vec 2.0. We anneal the temperature linearly from $\tau = 1.8$ to $\tau = 1.2$ over 25 epochs to gradually sharpen expert specialization. Entropy regularization is enabled at epoch 5 with $\lambda_{\text{ent}} = 10^{-4}$ to prevent premature gate collapse while allowing early specialization. Per-expert auxiliary weights are equalized to $[1.0, 1.0, 1.0]$.

Stage 3: Multi-domain fine-tuning (speech/singing: 50%/50%). Starting from the Stage 2 checkpoint, we fine-tune on a balanced mixture of speech and singing data (50/50 sampling per minibatch). All trainable components (gate, projections, auxiliary heads, ResNet encoders, and Wav2Vec CNN head) are updated end-to-end with a reduced learning rate of 3×10^{-5}. The Wav2Vec transformer backbone remains frozen. We use $\tau = 1.0$ and keep entropy regularization active ($\lambda_{\text{ent}} = 10^{-4}$). We limit to 5 epochs to avoid overfitting, as validation EER converges quickly after Stage 2.

5 Evaluation

5.1 Zero-Shot Cross-domain Evaluation

To motivate our approach, we first measure the zero-shot singing performance of the unified three-branch model of GenuVoice trained exclusively on speech datasets (ASVspoof 2019 and WaveFake). Table 2 shows the severity of the domain shift. While the model achieves strong detection on in-domain speech (0.38% EER on ASVspoof2019), it degrades to 43.16% EER on CtrSVDD, indicating a large cross-domain generalization failure. For completeness, we also report performance on SingFake* (53.63% EER). Note that the pooled singing score is dominated by CtrSVDD because it contains 12,326 of 12,464 samples (≈99%). We therefore emphasize per-dataset results in Table 2.

Table 3: Individual expert performance under speech-only training. All experts fail on singing data, confirming the domain shift is not expert-specific.

Expert Branch	Training	ASVspoof 2019	SingFake	CtrSVDD
Log-Mel (Spectral)	Speech only	1.71%	43.47%	43.84%
MFCC (Cepstral)	Speech only	1.68%	57.25%	46.44%
Wav2Vec 2.0 (Semantic)	Speech only	5.72%	52.90%	53.02%

Table 4: Individual expert performance after training on both speech and singing data. Log-Mel generalizes best to singing, while Wav2Vec transfers poorly.

Expert Branch	ASVspoof 2019	SingFake	CtrSVDD
Log-Mel (Spectral)	1.66%	**4.35%**	**6.07%**
MFCC (Cepstral)	1.64%	10.15%	9.39%
Wav2Vec 2.0 (Semantic)	5.66%	14.50%	20.53%

To diagnose this failure mode, we evaluate each expert branch independently. Table 3 reports individual expert performance under speech-only training, identical to the unified model in Table 2. All individual experts fail on singing (CtrSVDD EER: 43–53%), confirming the domain shift is not expert-specific. For comparison, Table 4 shows individual expert performance after training on both speech and singing data. The Log-Mel expert achieves 6.07% EER on CtrSVDD, substantially outperforming the speech-only baselines. However, our full fused model with multi-domain fine-tuning (Table 5) further reduces CtrSVDD EER to 1.82%, a 70% relative improvement over the best individual expert, demonstrating the benefit of multi-expert fusion under the speech-retentive training pipeline.

5.2 Multi-Domain Fine-Tuning Evaluation

We then evaluate GenuVoice, which is trained using our three-stage protocol with multi-domain fine-tuning, mixing speech and singing examples in a balanced 50%/50% ratio while retaining auxiliary losses for each expert. This strategy realigns the gate and enables cross-domain specialization while preserving speech performance. As shown in Table 5, multi-domain fine-tuning dramatically improves CtrSVDD performance from 43.16% to 1.82% EER while maintaining robust speech performance (0.38% EER on ASVspoof 2019). To evaluate generalization to unseen generators and codec conditions, we additionally test on ASVspoof 2021, which was held out entirely from training. GenuVoice achieves 8.89% EER on ASVspoof 2021 despite training only on ASVspoof 2019 and Wave-

Table 5: Performance after three-stage training with multi-domain fine-tuning. The model achieves strong gains on CtrSVDD (1.82% EER) while preserving robust speech performance (0.38% on ASVspoof 2019) and generalizing to unseen generators (8.89% on held-out ASVspoof 2021).

Dataset	Samples	In Training?	EER (%)	ROC-AUC
Target Domain (Singing)				
CtrSVDD	12,326	Yes (Stage 3)	**1.82**	0.9985
SingFake*	138	Yes (Stage 3)	44.20	0.5616
Source Domain (Speech)				
ASVspoof 2019	18,029	Yes	**0.38**	0.9991
ASVspoof 2021	145,331	No (held-out)	8.89	0.998
WaveFake	4,390	Yes	**0.36**	0.9999

Table 6: Baseline comparison: speech-trained detectors evaluated on singing benchmarks. All existing methods fail on CtrSVDD (37–62% EER); GenuVoice achieves 1.82%.

Method	CtrSVDD EER (%)	SingFake EER (%)
RawNet2 (Tak et al., 2021)	50.15	46.37
AASIST (Jung et al., 2022)	61.60	40.59
Wav2Vec2-AASIST (Tak et al., 2022)	37.24	51.45
Wav2Vec2-DF	44.91	51.45
Gohari et al. (Gohari et al., 2025)	11.72	–
GenuVoice (Ours)	**1.82**	**44.20**

Fake, demonstrating robustness to novel attack types and telephony compression absent from our training data.

Baseline comparison. To contextualize these results, we evaluate four established speech-trained detectors on singing benchmarks. As shown in Table 6, all methods achieve 37–62% EER on CtrSVDD—near or worse than random—confirming that speech-trained detectors fail catastrophically on singing. The most relevant prior work on singing deepfake detection (Gohari et al., 2025) reports a best EER of 11.72% on CtrSVDD. GenuVoice achieves 1.82% EER, representing a 95% relative improvement over the best speech-trained baseline (Wav2Vec2-AASIST at 37.24%) and an 84% relative improvement over the best prior singing-specific method. On SingFake, all methods including GenuVoice struggle (40–51% EER), consistent with the in-the-wild robustness challenges discussed below.

SingFake performance and gating miscalibration. While the proposed approach achieves strong performance on CtrSVDD (1.82% EER), performance on SingFake* remains challenging (44.20% EER). Beyond the data quality and distribution heterogeneity factors discussed above, the underperformance is most consistent with gating miscalibration under severe dataset imbalance. In Stage 3, the singing-domain batches are dominated by CtrSVDD (8,236 vs. 563 samples), so the gating network is optimized primarily for CtrSVDD's controlled conditions. The learned mixture weights do not transfer to SingFake's in-the-wild characteristics (web compression, residual accompaniment from source separation, and post-processing such as reverb and autotune), causing the fused model to over-weight experts in a configuration suboptimal for SingFake. In contrast, a single Log-Mel expert (Table 4) avoids this failure mode because it does not rely on routing/mixture weights. This behavior aligns with the known sensitivity of MoE routing to training distribution imbalance. Concrete remedies include rebalancing Stage 3 at the dataset level (e.g., equalizing CtrSVDD vs. SingFake sampling) and performing a lightweight gate calibration pass on SingFake while keeping expert encoders frozen.

Overall, the strong performance on CtrSVDD supports the claim that multi-domain fine-tuning with multi-expert fusion addresses the fundamental speech-to-singing domain shift when labeled data is reliable, while SingFake highlights a separate in-the-wild robustness challenge—driven by data scarcity and gating miscalibration—that remains critical future work for production deployment.

5.3 Adversarial Alignment Evaluation

To test whether adversarial domain alignment can reduce the speech-to-singing gap without explicit singing supervision, we evaluate Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2016). We systematically varied the singing data ratio (10%, 30%, 50%), training stages, and regularization options. Across all configurations, DANN yields substantially worse performance than our multi-domain fine-tuning approach. Table 7 summarizes the full sweep. Notably, even under an identical 50/50 mixing ratio—matching our main experiments—DANN achieves only 26.45% EER on CtrSVDD, compared to 1.82% for GenuVoice. This ensures the comparison is methodologically sound: under identical data conditions, supervised multi-domain fine-tuning dramatically outperforms adversarial alignment.

Why does DANN fail here? DANN promotes domain-invariant representations by adversarially aligning source and target features (Ganin et al., 2016). However, the most discriminative spoof cues in our setting are partially domain-specific:

Table 7: Systematic DANN evaluation. Even under matched 50/50 training conditions, DANN (26.45% EER) substantially underperforms GenuVoice (1.82%).

Configuration	Data Mix	Stage	CtrSVDD EER
DANN-1	10% singing	3	49.12%
DANN-2	30% singing	3	47.31%
DANN-3	30% + entropy reg	3	53.10%
DANN-4	30% unfrozen	3	52.33%
DANN-5	30% partial	2	53.63%
DANN-6	50% singing	3	26.45%
Best DANN	**50%**	**3**	**26.45%**
Ours (GenuVoice)	**50/50**	**3**	**1.82%**

Table 8: Ablation study on loss components. Each component contributes to improved singing detection while preserving speech performance.

Model Configuration	CtrSVDD EER (%)	Speech EER (%)
Baseline ($\mathcal{L}_{\text{task}}$ only)	12.34	0.45
+ Auxiliary loss ($\lambda_{\text{aux}} = 0.1$)	8.67	0.42
+ Entropy reg ($\lambda_{\text{ent}} = 10^{-4}$)	5.23	0.40
+ Diversity penalty ($\lambda_{\text{div}} = 0.1$)	3.45	0.39
Full model (all components)	**1.82**	**0.38**

prosodic or phonetic dynamics in speech versus harmonic texture and vibrato-related patterns in singing (Gohari et al., 2025). Enforcing invariance can therefore suppress task-relevant structure, yielding poor target performance—a pattern consistent with negative transfer (Wang et al., 2019; Kellerman, 1979).

5.4 Ablation Study

To validate our architectural and training design choices, we conduct systematic ablation studies on five key components. All ablations are performed using the three-stage protocol on CtrSVDD, with performance measured on the CtrSVDD evaluation split and ASVspoof 2019 speech test split.

5.4.1 Loss Component Analysis

Table 8 demonstrates that each loss component contributes meaningfully to final performance. The baseline model using only $\mathcal{L}_{\text{task}}$ achieves 12.34% EER on CtrSVDD. Adding auxiliary expert losses ($\lambda_{\text{aux}} = 0.1$) improves to 8.67% EER, entropy regularization ($\lambda_{\text{ent}} = 10^{-4}$) further reduces to 5.23%, and diversity penalty ($\lambda_{\text{div}} = 0.1$) yields 3.45%. The full model combining all components achieves 1.82% EER, representing an 85% relative improvement over the baseline while maintaining robust speech performance (0.38% EER).

Table 9: Effect of entropy regularization λ_{ent} on gate collapse. Optimal value balances expert utilization without over-smoothing.

λ_{ent}	CtrSVDD EER (%)	Gate Collapse?	Avg $\alpha_{\max}$	Notes
0.0	3.45	Yes	0.92	Collapses to Wav2Vec
10^{-5}	2.67	Partial	0.78	Still biased
$\mathbf{10^{-4}}$ (ours)	**1.82**	No	0.52	Balanced
5×10^{-4}	2.34	No	0.42	Too uniform
10^{-3}	4.12	No	0.38	Over-smoothed

Table 10: Effect of diversity penalty λ_{div} on expert specialization. Optimal value encourages complementary representations.

λ_{div}	CtrSVDD EER (%)	Avg $\cos(\mathbf{z}_i, \mathbf{z}_j)$	Notes
0.0	3.89	0.67	Experts similar
0.05	2.54	0.45	Moderate diversity
0.1 (ours)	**1.82**	0.23	Good diversity
0.2	2.12	0.15	Over-penalized
0.5	3.45	0.08	Experts too different

5.4.2 Entropy Regularization Strength

Table 9 investigates the effect of entropy regularization λ_{ent} on preventing gate collapse. Without regularization ($\lambda_{\text{ent}} = 0$), the gate collapses to a single expert (average $\alpha_{\max} = 0.92$), yielding 3.45% EER. Very weak regularization ($\lambda_{\text{ent}} = 10^{-5}$) partially mitigates collapse but remains biased. Our chosen value ($\lambda_{\text{ent}} = 10^{-4}$) achieves balanced expert utilization ($\alpha_{\max} = 0.52$) and best performance (1.82% EER). Stronger regularization oversmooths the gate, degrading performance.

5.4.3 Diversity Penalty Strength

Table 10 evaluates the impact of diversity penalty λ_{div} on expert specialization. Without diversity penalty ($\lambda_{\text{div}} = 0$), experts learn similar representations (average cosine similarity $= 0.67$), resulting in 3.89% EER. Our chosen value ($\lambda_{\text{div}} = 0.1$) encourages sufficient diversity (cosine similarity $= 0.23$) for complementary expertise, achieving 1.82% EER. We find that a large penalty ($\lambda_{\text{div}} = 0.5$) forces experts too far apart, degrading performance to 3.45%.

5.4.4 Temperature Annealing Schedule

Table 11 evaluates different temperature schedules during Stage 2. Fixed temperature ($\tau = 1.0$ or $\tau = 1.2$) yields suboptimal specialization (EER $\geq 3.87\%$). Our annealing schedule ($\tau : 1.8 \rightarrow 1.2$) achieves best performance (1.82% EER) by initially allowing broad exploration then gradually sharpening expert selection. More aggressive annealing ($\tau : 2.0 \rightarrow 1.0$ or $\tau : 1.8 \rightarrow 0.8$) either converges too quickly or over-specializes, degrading performance.

Table 11: Effect of temperature annealing schedule in Stage 2. Gradual annealing from 1.8 to 1.2 achieves optimal expert specialization.

Schedule	CtrSVDD EER (%)	Speech EER (%)	Notes
Fixed $\tau = 1.0$	4.56	0.40	No specialization
Fixed $\tau = 1.2$	3.87	0.39	Weak specialization
$\tau : \mathbf{1.8} \rightarrow \mathbf{1.2}$ (ours)	**1.82**	**0.38**	Best performance
$\tau : 2.0 \rightarrow 1.0$	2.34	0.38	Too aggressive
$\tau : 1.8 \rightarrow 0.8$	3.12	0.41	Over-specialization

Table 12: Effect of speech-singing mixing ratio in Stage 3. Balanced 50:50 ratio achieves best cross-domain performance.

Ratio (Speech:Singing)	CtrSVDD EER (%)	Speech EER (%)	Notes
90:10	6.78	0.36	Insufficient singing
70:30	3.45	0.37	Better adaptation
50:50 (ours)	**1.82**	**0.38**	Best balance
30:70	2.67	0.42	Speech degradation
10:90	4.23	0.51	Catastrophic forgetting

5.4.5 Speech-Singing Mixing Ratio

We then investigate the effect of speech-singing mixing ratio during Stage 3 fine-tuning. As shown in Table 12, insufficient singing data (90:10 ratio) prevents effective adaptation (6.78% EER). Our balanced 50:50 ratio achieves optimal performance (1.82% EER on CtrSVDD, 0.38% on speech), successfully adapting to singing while preserving speech capabilities. Singing-heavy ratios (30:70 or 10:90) cause catastrophic forgetting of speech patterns, degrading speech EER to 0.42-0.51% while providing limited benefit on singing.

6 Discussion

Why multi-domain fine-tuning beats adversarial invariance? A central question in cross-domain audio forensics is whether to enforce feature *invariance* (via adversarial alignment) or to learn *domain-aware* decision boundaries under supervision. As shown in the systematic DANN sweep (Table 7), even under identical 50/50 data conditions, DANN-based adversarial alignment yields 26.45% EER on CtrSVDD, whereas our multi-domain fine-tuning achieves 1.82% EER while preserving speech performance (0.38% EER on ASVspoof 2019). This confirms that the improvement is due to the training strategy rather than data imbalance.

Why does adversarial alignment fail? DANN and related methods are most effective when the domain shift is primarily a *nuisance variable* (e.g., microphone/channel mismatch, compression, background noise, or recording environment) that does not overlap with the discriminative cues needed for spoof detection. In such cases, encouraging invariance can remove spurious domain indicators while preserving label-relevant structure. The DANN results reflect **negative transfer** in our cross-domain deepfake detection setting. DANN encourages speech and singing features to be aligned. However, spoof cues are partially domain-dependent. Speech deepfakes often manifest through prosodic/phonetic inconsistencies, while singing introduces distinct acoustic structure (e.g., sustained harmonics, vibrato, and pitch-range dynamics) (Gohari et al., 2025). Enforcing invariance can therefore suppress task-relevant structure and yield poor target performance, consistent with our systematic sweep in Table 7. This pattern aligns with negative-transfer observations in transfer learning when alignment objectives erase label-relevant features (Wang et al., 2019; Kellerman, 1979).

Multimodel sensing-assisted deepfake audio detection. The future extensions of GenuVoice can incorporate acoustic-sensing modules that model source-production and propagation cues, for example, through human-acoustics descriptors (Wang et al., 2023a), device-response signatures, or other sensor-inspired features. Such physically grounded cues would be complementary to the current spectral, cepstral, and self-supervised branches, and may improve robustness when generative models mimic surface-level content but still fail to reproduce the full acoustic behavior of real recordings.

7 Conclusion

In this work, we propose the first unified framework for joint speech and singing deepfake detection. We show that speech-trained deepfake detectors fail under speech-to-singing domain shift, with existing methods achieving 37–62% EER on CtrSVDD despite strong speech accuracy, indicating that singing is a structurally distinct acoustic domain. We propose GenuVoice, a unified mixture-of-experts framework that adaptively fuses spectral, cepstral, and semantic representations while enforcing expert specialization through auxiliary supervision and speech-retentive multi-domain fine-tuning. GenuVoice achieves 1.82% EER on CtrSVDD, a 95% relative improvement over the best speech-trained baseline, while preserving speech performance (0.38% EER on ASVspoof 2019) and generalizing to unseen generators (8.89% EER on held-out ASVspoof 2021). In the future, we expect new cues to complement learned representations for more robust deepfake detection.

Limitations

Gate interpretability and expert accountability. Although the gate provides adaptive expert weighting, mixture decisions can remain opaque. We can infer behavior indirectly (e.g., improved singing detection is consistent with upweighting spectral experts), but we do not provide a systematic interpretability analysis. Better understanding when and why the gate favors an expert could enable debugging of failure cases and increase trust in deployment. Future work could include gate statistics by dataset, confidence calibration, and contribution analysis under controlled perturbations.

Limited coverage of generators and post-processing. While we demonstrate generalization to unseen generators via held-out ASVspoof 2021 (8.89% EER), evaluation remains restricted to the generators and post-processing pipelines represented in the chosen benchmarks. Real-world deepfakes may involve novel vocoders or heavy editing that shifts artifacts beyond what current benchmarks capture. Further robustness testing using leave-one-generator-out protocols or newer corpora remains important future work.

Computational cost of multi-branch inference. Multiple experts increase inference latency and memory footprint relative to a single-branch model. While this design is motivated by cross-domain robustness, it may be costly for real-time or edge deployment. Practical deployment may require distillation, expert pruning, or conditional computation where only a subset of experts is evaluated for each input.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. 2019. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*.

Joel Frank and Lea Schönherr. 2021. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Mahyar Gohari, Davide Salvi, Paolo Bestagini, and Nicola Adami. 2025. Audio features investigation for singing voice deepfake detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hanqing Guo, Guangjing Wang, Yuanda Wang, Bocheng Chen, Qiben Yan, and Li Xiao. 2023. Phantomsound: Black-box, query-efficient audio adversarial attack via split-second phoneme injection. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, pages 366–380.

Yunqi Hao, Yihao Chen, Minqiang Xu, Jianbo Zhan, Liang He, Lei Fang, Sian Fang, and Lin Liu. 2025. Wav2df-tsl: Two-stage learning with efficient pre-training and hierarchical experts fusion for robust audio deepfake detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Michael I Jordan and Robert A Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

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 *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6367–6371. IEEE.

Eric Kellerman. 1979. Transfer and non-transfer: Where we are now. *Studies in second language acquisition*, 2(1):37–57.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kavita Kumari, Maryam Abbasihafshejani, Alessandro Pegoraro, Phillip Rieger, Kamyar Arshi, Murtuza Jadliwala, and Ahmad-Reza Sadeghi. 2025. Voiceradar: Voice deepfake detection using micro-frequency and compositional analysis. In *NDSS*.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and 1 others. 2023. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*.

Viola Negroni, Davide Salvi, Alessandro Ilic Mezza, Paolo Bestagini, and Stefano Tubaro. 2025. Leveraging mixture of experts for improved speech deepfake detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jeeweon 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*.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*.

Guangjing Wang, Qiben Yan, Shane Patrarungrong, Juexing Wang, and Huacheng Zeng. 2023a. Facer: Contrastive attention based expression recognition via smartphone earpiece speaker. In *IEEE INFOCOM 2023-IEEE conference on computer communications*, pages 1–10. IEEE.

Yuanda Wang, Bocheng Chen, Hanqing Guo, Guangjing Wang, Weikang Ding, and Qiben Yan. 2025. Clearmask: Noise-free and naturalness-preserving protection against voice deepfake attacks. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, pages 696–709.

Yuanda Wang, Hanqing Guo, Guangjing Wang, Bocheng Chen, and Qiben Yan. 2023b. Vsmask: Defending against voice synthesis attack via real-time predictive perturbation. In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pages 239–250.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302.

Yongyi Zang, Jiatong Shi, You Zhang, Ryuichi Yamamoto, Jionghao Han, Yuxun Tang, Shengyuan Xu, Wenxiao Zhao, Jing Guo, Tomoki Toda, and 1 others. 2024a. Ctrsvdd: A benchmark dataset and baseline analysis for controlled singing voice deepfake detection. *arXiv preprint arXiv:2406.02438*.

Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. 2024b. Singfake: Singing voice deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12156–12160. IEEE.

Yongmao Zhang, Jian Cong, Heyang Xue, Lei Xie, Pengcheng Zhu, and Mengxiao Bi. 2022. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7237–7241. IEEE.