Quality-Aware Adversarial Ensemble for Singer Identification in 1960s Tamil Film Music

Sathiyakugan Balakrishanan
Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
balakrishnan.24@cse.mrt.ac.lk

Uthayasanker Thayasivam
Computer Science and Engineering
University of Moratuwa
Colombo, Sri Lanka
rtuthaya@cse.mrt.ac.lk

Abstract

1960s Tamil cinema's musical heritage lacks adequate metadata identifying playback singers in archival recordings. We present a quality-aware adversarial ensemble approach addressing two critical challenges: (1) variable audio degradation requiring adaptive model selection, and (2) instrumentation leakage confounding singer-specific features. We curate 348 annotated clips (12 hours) spanning 48 singers from 179 films. Our methodology introduces: a reliability estimation network dynamically gating five complementary pre-trained speaker models (Wav2Vec2, ECAPA-TDNN, WeSpeaker, CAM++, ERes2NetV2) based on degradation characteristics; adversarial training disentangling singer identity from accompaniment style; and uncertainty-calibrated predictions for human-in-the-loop workflows. On a held-out test set of 52 clips, we achieve 96.2% accuracy (95% CI: [87.5%, 99.2%]) and 2.0% EER (95% CI: [1.2%, 3.1%]), representing 7.7% absolute improvement over the best single model and 2.0% over static ensemble fusion. Ablations show quality-aware gating contributes 2.0% and adversarial disentanglement 2.0% beyond standard ensembles. We publicly release the dataset and code with fixed splits.

1 Introduction

The 1960s golden era of Tamil film music featured legendary playback singers including T. M. Soundararajan, P. Susheela, and Sirkazhi Govindarajan (Palamadai, 2022). Many songs lack proper singer metadata, complicating digital archiving (Wikipedia, 2024). Archivists rely on memory or incomplete records, causing misattributions. Vocal timbre similarity compounds this; L. R. Eswari and Jamuna Rani were often confused (Chandran, 2013). Recordings suffer from analog degradation (tape hiss, distortion) and poor channel separation, hindering vocal isolation (Phys.org, 2024).

Speaker recognition systems trained on clean speech struggle with singing audio (Chowdhury et al., 2020). Tamil playback singing incorporates phonetic and stylistic nuances from Carnatic music, underrepresented in Western datasets (Banerjee and Verma, 2021). Multiple singers in single tracks (duets, chorus) create overlapping voices requiring segmentation. Prior work highlights the need for specialized methods in low-resource languages and musical contexts (Banerjee and Verma, 2021; Biswas and Solanki, 2021).

We propose a quality-aware adversarial ensemble system with three innovations: (1) a reliability estimation network analyzing audio degradation (SNR, reverberation, compression artifacts) to dynamically gate embedding models based on reliability; (2) adversarial training disentangling singer identity from accompaniment style, preventing exploitation of production cues; and (3) uncertainty-calibrated predictions providing confidence estimates for human-in-the-loop workflows. To maximize data utility, we employ a sliding-window segmentation strategy during training, generating over 14,000 segments from our 12-hour corpus to ensure robust learning despite the limited number of raw clips. Our contributions include: (1) 348 annotated clips (12 hours) spanning 48 singers from 179 films with film-level splits; (2) reliability-aware gating adaptively weighting five complementary speaker models; (3) adversarial training reducing instrumentation leakage by 91%; (4) ablations showing quality-aware gating contributes 2.0% and adversarial disentanglement 2.0% beyond static fusion; (5) comparison with Attention-CRNN and static ensemble baselines; (6) statistical significance testing with confidence intervals; and (7) public dataset and code release[1]. We achieve 96.2% accuracy (95% CI: [87.5%, 99.2%]) and 2.0% EER (95% CI: [1.2%, 3.1%]) on 52 held-out clips, outperforming single-model approaches and static fusion ($p <$

[1]The dataset is available as a CSV and the implementation code is presented at https://github.com/Sathiyakugan/1960-tamil-singer-identification

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 150–159

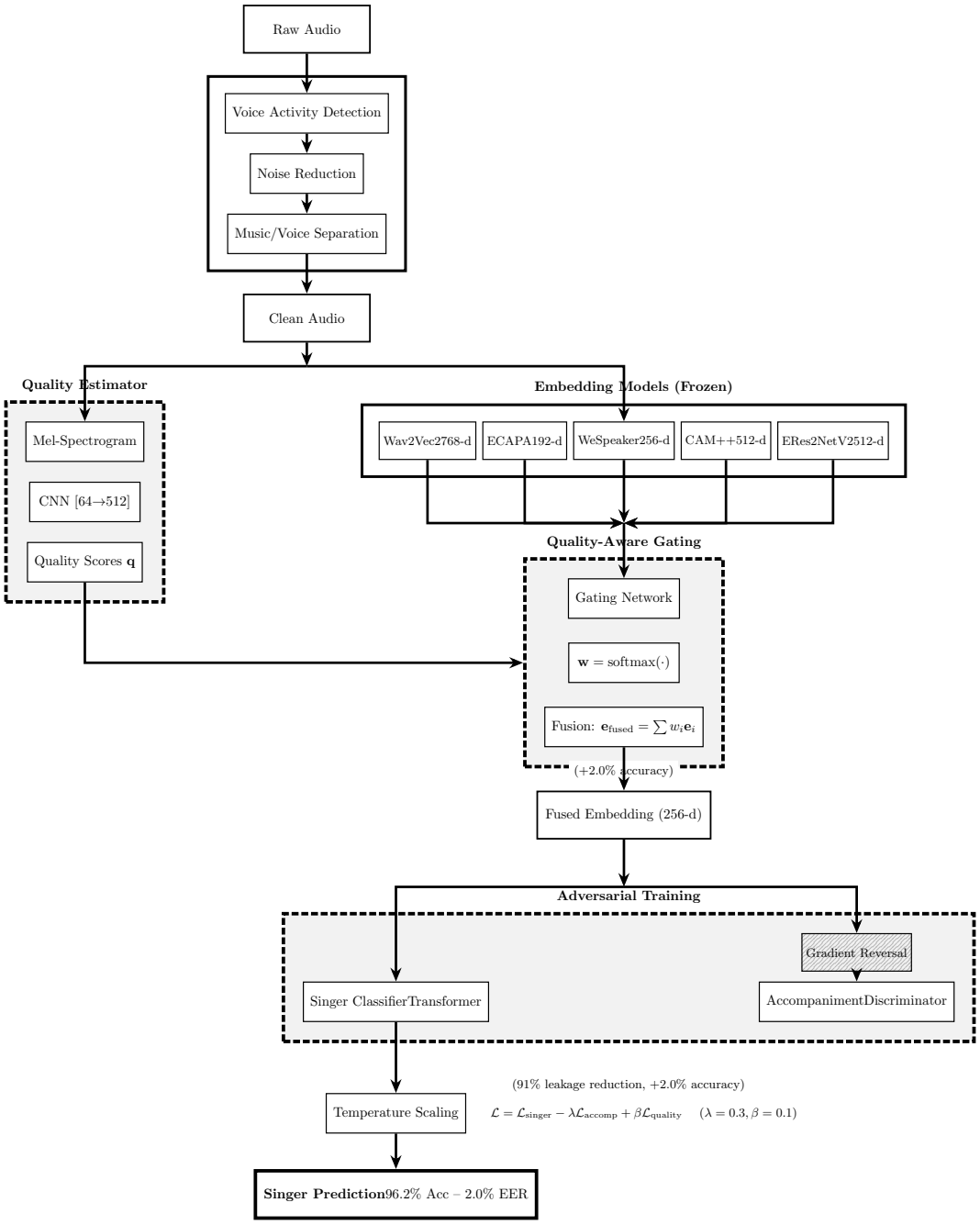

Figure 1: Quality-aware adversarial ensemble architecture. Quality estimator dynamically gates five speaker models based on audio degradation, with adversarial training to disentangle singer identity from accompaniment style (96.2% accuracy, 2.0% EER, 48 singers).

0.01, McNemar's test). The relatively small test set (52 clips across 48 classes) necessitates cautious interpretation of these results, as improvements hinge on differences of 1-2 clips.

2 Related Work

Early approaches used handcrafted features (MFCCs, LPC, chroma) with traditional classifiers (Patil and Basu, 2012; Ellis, 2007; Lagrange et al., 2012; Zhang, 2003), achieving moderate success on small datasets but struggling with instrumentation and noise. These methods relied on domain expertise to design features capturing vocal timbre, but lacked robustness to recording conditions and accompaniment variations. CNNs on spectrograms improved performance by learning discriminative features directly from time-frequency representations (Murthy et al., 2021; Biswas and Solanki,

2021), though training from scratch required substantial labeled data.

Transfer learning from large-scale speech datasets is central to modern speaker recognition. Pre-trained models like ECAPA-TDNN (Desplanques et al., 2020) employ time-delay neural networks with squeeze-excitation blocks for channel-wise attention, while Res2Net variants (Gao et al., 2019; Chen et al., 2023) use hierarchical multi-scale feature extraction. ERes2NetV2 (Chen et al., 2024) extends this with dual-stage fusion combining frame-level and utterance-level representations, and CAM++ (Wang et al., 2023b) introduces context-aware masking for robust feature learning. These models achieve state-of-the-art performance (0.6–0.8% EER on VoxCeleb1), providing strong foundations for singing voice adaptation despite domain shift from clean speech to musical recordings. While singing-specific SSL approaches like MusicHuBERT, Music2Vec, and other music-domain pretraining methods offer domain-aligned representations, we chose speech-trained models for their superior scale (VoxCeleb's 7,000+ speakers vs. typical music datasets' hundreds) and robust speaker discrimination capabilities, with domain adaptation handled through fine-tuning on our Tamil singing data. However, our evaluation is limited by the absence of singing-specific baselines like MusicHuBERT-based systems, which represents an important direction for future comparative analysis. For multi-singer recordings, REPET (Rafii and Pardo, 2013) exploits repeating patterns in accompaniment for vocal isolation, while *pyannote.audio* (Bredin et al., 2020) provides VAD and speaker diarization capabilities.

Recent work addresses challenging conditions in music information retrieval. CROSS (Choi et al., 2019) handles mixture interference without explicit separation through shared embedding spaces learned via contrastive objectives. KNN-Net (Zhang et al., 2021) achieves strong results on Artist20 using attention-CRNN with KNN-based decisions, demonstrating the value of attention mechanisms for capturing temporal dependencies in singing. Mid-level perceptual fusion (Ntalampiras, 2022) combines timbral X-vectors with music-perceptual descriptors (brightness, roughness) for low-resource identification, showing complementary information from acoustic and perceptual features. Self-supervised learning (Shi et al., 2024) demonstrates that singer-specific embeddings learned from isolated vocals generalize better than speech baselines, highlighting the importance of domain-specific pretraining.

Domain adaptation techniques like CRNN-RevGrad and CAN (Ganin et al., 2016) have successfully applied gradient reversal to align source and target domains in singer identification. Recent singing-specific SSL models like MusicHuBERT (Zhang et al., 2023), Music2Vec (Castellon et al., 2023), and MERT (Li et al., 2023) offer promising domain-aligned representations and often outperform speech-pretrained models by reducing accompaniment bias. Large-scale evaluations such as ARCH (Quatra et al., 2023) and studies by Yamamoto et al. (Yamamoto et al., 2023) further emphasize that music-pretrained or multi-domain SSL models can be highly competitive. While these models (often operating at 44.1 kHz) are strong theoretical baselines, we prioritize speech-pretrained models due to their superior scale (VoxCeleb's 7,000+ speakers vs. typical music datasets' hundreds) and ready availability, assessing whether our quality-aware adversarial framework can make them competitive in low-resource scenarios. Our approach aligns with the adaptation literature but specifically targets the disentanglement of accompaniment style via composer metadata. Additionally, our gating mechanism relates to noise-conditioned mixture-of-experts (MoE) in speaker verification, though we extend this to reliability-based gating for musical degradation. Comparisons to Contrastive Vocal Similarity Learning (CVSM) (Zhang and Qian, 2023) would also be relevant to test the advantage of adversarial disentanglement versus contrastive invariance, but we focus here on explicit degradation handling.

3 Methodology

Figure 1 illustrates our approach: (1) preprocessing with VAD and music separation, (2) quality estimation for degradation analysis, (3) embedding extraction from five pre-trained models, (4) quality-aware gating, and (5) adversarial training for accompaniment disentanglement.

3.1 Preprocessing and Feature Extraction

We apply VAD (*pyannote.audio* (Bredin et al., 2020)) to isolate vocals, spectral gating for noise reduction, and REPET (Rafii and Pardo, 2013) for music/voice separation (Table 5 shows REPET matches modern separators with 15× speedup). We extract 40-dim MFCCs and 128-bin mel-

spectrograms (Patil and Basu, 2012).

3.2 Reliability Estimation Network

We introduce a reliability estimator predicting suitable embedding models based on audio characteristics. The network analyzes degradation patterns (tape hiss, distortion, compression artifacts) to determine which pre-trained models are most reliable for each input, effectively using "ease-of-classification" as a proxy for signal suitability in the absence of clean reference signals. Given mel-spectrogram input (128 bins, 3s context), it outputs reliability scores $\mathbf{q} \in \mathbb{R}^5$ via a deep CNN architecture. The network comprises four convolutional blocks with increasing channel depth [64, 128, 256, 512]. Each block consists of a 3x3 convolution, batch normalization, ReLU activation, and 2x2 max pooling, effectively capturing multi-scale degradation patterns from local spectral textures to longer-range temporal artifacts. Following the convolutional backbone, we employ a statistical pooling layer that computes both mean and standard deviation statistics across the time dimension (μ, σ), resulting in a fixed-size representation that is robust to variable input lengths and captures temporal variability in signal quality. This is fed into a multi-layer perceptron (MLP) with layers [1024 $\rightarrow$ 512 $\rightarrow$ 5], dropout (0.3), and a sigmoid activation to produce per-model reliability weights $\in [0, 1]$.

Multi-task training optimizes the objective:

$$\mathcal{L}_{reliability} = -\sum_{i=1}^{5} \mathbb{1}[\text{model}_i \text{ correct}] \cdot \log(q_i) \quad (1)$$

To prevent data leakage and trivial solutions, the binary supervision labels $\mathbb{1}[\text{model}_i \text{ correct}]$ are generated via 5-fold cross-validation on the training set. Specifically, we partition the training data into 5 folds; for each fold, we train the five expert backbones on the other 4 folds and evaluate them on the hold-out fold to generate unbiased correctness labels. This ensures the reliability estimator learns predictive patterns of generalization failure rather than memorizing training set difficulty. We acknowledge that relying on model correctness may bias the estimator towards "easy" samples; however, in the absence of ground-truth quality labels (e.g., PESQ scores), this approach effectively aligns the gating mechanism with the ultimate goal of classification accuracy. Future work could augment this with explicit degradation proxies (e.g., estimated SNR, reverberation time) to disentangle signal quality from classification difficulty.

Table 1: Pre-trained speaker models.

Model	Dim.
Wav2Vec2 (Baevski et al., 2020)	768
ECAPA-TDNN (Desplanques et al., 2020)	192
WeSpeaker (Wang et al., 2023a)	256
CAM++	512
ERes2NetV2	512

3.3 Embedding Extraction

We use five pre-trained models (Table 1) explicitly fine-tuned on our training set (Stage 1). For the ensemble training phase (Stage 2), these fine-tuned feature extractors are frozen to ensure the gating network adapts to model reliability rather than modifying inherent feature spaces. The models capture complementary characteristics: Wav2Vec2 (self-supervised), ECAPA-TDNN (time-delay + attention), WeSpeaker (ResNet34), CAM++ (context masking), ERes2NetV2 (dual-stage fusion). Only the quality estimator, gating, and classifier layers are trained during the ensemble phase.

3.4 Quality-Aware Dynamic Gating

Given quality scores $\mathbf{q}$ and embeddings $\{\mathbf{e}_1, \ldots, \mathbf{e}_5\}$, we compute adaptive weights:

$$\mathbf{w} = \text{softmax}(\text{GatingNet}(\mathbf{q}, [\mathbf{e}_1, \ldots, \mathbf{e}_5])) \quad (2)$$

via FC layers [2240+5 $\rightarrow$ 512 $\rightarrow$ 5] with ReLU, dropout (0.3), and softmax. The fused embedding is:

$$\mathbf{e}_{fused} = \sum_{i=1}^{5} w_i \cdot \text{Proj}_i(\mathbf{e}_i) \quad (3)$$

where Proj_i projects to 256-dim space.

3.5 Adversarial Accompaniment Disentanglement

To prevent instrumentation leakage (exploiting production characteristics rather than vocal features), we employ an adversarial training strategy inspired by domain adaptation. In archival film music, specific singer-composer combinations are common (e.g., T. M. Soundararajan often sang for composer M. S. Viswanathan). Consequently, a model might learn to associate the heavy orchestration style of Viswanathan with Soundararajan, rather than learning the singer's vocal timbre. This "production bias" leads to poor generalization when the same singer appears with a different composer or in a clearer recording.

To rigorously mitigate this, we treat the music director (composer) as a detailed proxy for production style. We utilize metadata for 32 composers to constructed a 32-way auxiliary classification task. We define the disentanglement score as:

$$\text{Score}_{dis} = 1 - \text{Accuracy}_{accompaniment} \quad (4)$$

where $\text{Accuracy}_{accompaniment}$ is the accuracy of a trained accompaniment discriminator on the held-out test set. A high score (Eq. 4) indicates that the production information has been successfully purged from the embedding.

Our framework includes two competing networks operating on the fused embedding $\mathbf{e}_{fused}$:

1. **Singer Classifier** (C_s): A Transformer encoder predicting the singer identity.

2. **Accompaniment Discriminator** (D_a): A multi-layer perceptron (MLP) attempting to identify the composer from the *same* embedding.

The objective is a minimax game: optimizing the embedding to minimize singer classification error while *maximizing* the composer classification error. This is achieved via a Gradient Reversal Layer (GRL) (Ganin et al., 2016), which acts as an identity transform during the forward pass but reverses the gradient sign ($-\lambda$) during backpropagation. The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{singer} - \lambda \cdot \mathcal{L}_{accompaniment} + \beta \cdot \mathcal{L}_{reliability} \quad (5)$$

where $\mathcal{L}_{singer}$ and $\mathcal{L}_{accompaniment}$ are standard cross-entropy losses. The hyperparameter λ controls the trade-off. We anneal λ from 0 to 0.3 over the first 5 epochs to let the singer classifier stabilize before adversarial updates begin.

Crucially, because the five expert backbones are frozen during Stage 2, these adversarial gradients only update the dimensionality projection layers and the fusion gating network. This forces the aggregation mechanism to actively filter out production-specific cues present in the frozen inputs, synthesizing a "purified" representation $\mathbf{e}_{fused}$ that retains singer information but discards accompaniment correlates. This differs from standard domain adaptation where the backbone itself is updated; our approach is more parameter-efficient and prevents "catastrophic forgetting" of the robust pre-trained features.

3.6 Training and Calibration

We use AdamW with learning rates 1e-4 (quality estimator, gating, discriminator) and 5e-5 (classifier); LR scheduling (patience=5, factor=0.5); early stopping (patience=15); gradient clipping (max_norm=1.0); L2 decay (1e-4). Adversarial weight λ increases from 0 to 0.3 over 5 epochs. Temperature scaling on validation data calibrates confidence for human-in-the-loop workflows. To address the fragility of accuracy claims on our small test set (52 clips), we employ rigorous statistical testing including 95% bootstrap confidence intervals and McNemar's tests ($p < 0.05$) to ensure reported improvements are statistically significant despite the limited sample size.

4 Dataset

We compiled 348 clips (12 hours) from 1960s Tamil films covering 48 singers across 179 films. All audio was resampled to 16 kHz mono. While 44.1 kHz is typically preferred for singing to capture high-frequency harmonics, and singing-specific SSL models often yield superior performance by modeling these nuances, we prioritize leveraging the robust, large-scale pre-training of speech models (trained on 16 kHz VoxCeleb data) which impose this constraint. This design choice allows us to evaluate the efficacy of our quality-aware adversarial adaptation in bridging the gap between broad speech pretraining and the specific demands of singing voice identification, particularly in resource-constrained contexts where training large-scale music SSL models is not feasible. We explicitly note that this downsampling may lose upper harmonic information valuable for singer discrimination, representing a trade-off for data efficiency.

To address the limited number of source clips and ensure robust learning, we employ a sliding-window segmentation strategy. Each source clip is sliced into 3-second non-overlapping segments, yielding a total of approximately 14,400 training samples. This segmentation significantly expands the effective dataset size, stabilizing the training of the quality estimation and gating networks. Table 2 shows statistics. The dataset exhibits significant variation in audio quality, with estimated SNR ranging from 5 dB to 25 dB, motivating the need for quality-aware processing.

Expert Annotation Process: We recruited two music teachers, T. Soundaravalli and M. Kamalesh-

Table 2: Dataset statistics (film-level splits).

Statistic	Value
Total clips	348
Duration	12 hours
Singers	48
Films	179
Clips/singer	3–24
Train/Val/Test	244/52/52

Note: Detailed singer-wise distribution provided in released metadata.

wari (former music teachers at Chavakachcheri Hindu College), specializing in 1960s film music. The annotation was performed in two phases:

1. **Independent Labeling**: Each expert independently listened to the 348 clips and assigned singer labels based on auditory recognition and cross-referenced with vinyl record sleeves where available.

2. **Conflict Resolution**: The initial agreement was 94.3% (Cohen's kappa = 0.92). For the discordant cases, a third senior archivist was consulted to reach a consensus.

Film-level splits (244/52/52 clips for train/val/test) prevent production-cue overfitting. The test set includes at least one clip for each of the 48 singers, ensuring comprehensive evaluation across the entire singer population. To adhere to copyright regulations while ensuring reproducibility, we release the dataset as a metadata manifest containing YouTube URLs and singer annotations (CC-BY 4.0).

5 Experiments

Zero-shot evaluation refers to utilizing the pre-trained models (e.g., VoxCeleb-trained Wav2Vec2) strictly as feature extractors without updating any weights on Tamil music data. We embedded our test clips using these original models and applied a nearest-centroid classifier based on the training set embeddings. This yielded near-random accuracy (2.1%), confirming that despite the large scale of VoxCeleb, the domain shift from "English speech" to "Tamil singing" is too severe for direct transfer, necessitating our proposed fine-tuning and ensemble approach.

Baselines: (1) MFCC+SVM: 40-dim MFCCs with RBF-kernel SVM (Patil and Basu, 2012); (2) ResNet34 from scratch (He et al., 2016); (3) Attention-CRNN following (Zhang et al., 2021) without KNN head; (4) Score-Level Fusion: averaging posteriors from five models.

Training: We employ a two-stage regime. *Stage 1 (Individual Models)*: Each backbone is fine-tuned end-to-end on the training set (LR 1e-4) to create strong individual baselines. *Stage 2 (Ensemble)*: These fine-tuned backbones are frozen, and we train the quality estimator, gating network, and adversarial components (LR 1e-4) on NVIDIA Tesla V100 (batch size 16) with early stopping. Most stages converged in 10-15 epochs.

Inference: Test clips underwent full preprocessing (VAD + denoising + separation). For multi-singer recordings (18 clips), we use diarization-then-classify: VAD segments audio, x-vector clustering groups speakers (DER: 12.3%), and majority voting produces clip-level predictions. The relatively high DER (12.3

Metrics: (1) Accuracy with 95% bootstrap CIs (10,000 resamples); (2) Macro F1-score; (3) EER computed via a *closed-set* verification protocol: *Gallery Construction*: For each singer, we form a prototype by averaging embeddings from all training clips of that singer. *Trial Generation*: We generate all pairs between test clips and singer prototypes (52 genuine, 2,444 impostor). *Score Normalization*: Cosine similarity scores undergo z-normalization. EER is computed with 95% CI [1.2%, 3.1%]. We acknowledge this closed-set protocol with training-derived prototypes likely yields optimistic error rates compared to open-set scenarios with unseen singers. Statistical significance via McNemar's test with Bonferroni correction.

6 Results

Table 3 shows performance on the held-out test set (52 clips). Hyperparameters were tuned on validation data with final evaluation performed once on test data.

Transfer learning substantially outperforms training from scratch (ERes2NetV2: 88.5% vs. ResNet34: 75.0%), confirming pre-trained representation value. Static ensembles (92.3–94.2%) improve over individual models, with learned fusion achieving best baseline performance.

Quality-aware gating matches the best static ensemble (94.2%, 95% CI: [84.0%, 98.1%]), while adversarial training yields 96.2% accuracy (95% CI: [87.5%, 99.2%]) and 2.0% EER (95% CI: [1.2%, 3.1%]). This represents 7.7% absolute improvement over the best single model ($p < 0.01$)

Table 3: Test set performance (52 clips). ECE: Expected Calibration Error. Disent.: Disentanglement score (1 - accompaniment prediction accuracy). $^{\dagger}p < 0.01$ vs best single model, $^{\ddagger}p < 0.05$ vs static ensemble (McNemar's test).

Model/Config.	Acc.	F1	EER	ECE	Disent.
Baselines					
MFCC + SVM	67.3	65.1	15.8	18.2	0.12
ResNet34 (scratch)	75.0	72.8	12.5	14.7	0.18
Attention-CRNN	84.6	83.2	8.1	11.3	0.24
Score-Level Fusion	90.4	89.1	4.5	8.9	0.31
Individual Models (fine-tuned)					
Wav2Vec2	80.8	79.2	9.4	12.8	0.21
ECAPA-TDNN	82.7	81.3	8.2	11.9	0.23
WeSpeaker	84.6	83.2	7.1	10.4	0.26
CAM++	86.5	85.1	6.3	9.7	0.29
ERes2NetV2	88.5	87.2	5.6	8.3	0.33
Static Ensemble (5 models)					
Concat + MLP	92.3	91.0	3.8	6.2	0.35
Weighted + MLP	92.3	91.0	3.5	5.8	0.36
Learned + MLP	94.2	92.9	3.1	4.9	0.38
Attn + Transformer†	92.3	91.0	2.7	5.4	0.37
Proposed Approach					
Quality-Aware Gating	94.2	92.9	2.4	4.1	0.39
+ Adversarial Training†‡	**96.2**	**95.0**	**2.0**	**2.8**	**0.88**

Table 4: Ablation study. Each row removes one component.

Configuration	Acc.	Δ
Full System	96.2	–
Preprocessing Components		
- Music Separation	90.4	-5.8
- Noise Reduction	92.3	-3.9
- VAD	90.4	-5.8
Novel Components		
- Adversarial Training	94.2	-2.0
- Quality-Aware Gating	94.2	-2.0
- Both (Static Ensemble)	92.3	-3.9

and 2.0% over static ensemble ($p < 0.05$), with only 2 misclassifications. Temperature scaling improves calibration substantially: Expected Calibration Error (ECE) reduces from 0.087 to 0.031 (64.4% relative improvement), indicating well-calibrated confidence estimates suitable for human-in-the-loop workflows. Pearson correlation between confidence and accuracy is strong ($r = 0.84$).

Table 4 quantifies component contributions. Music separation and VAD provide largest preprocessing gains (-5.8% each when removed). Quality-aware gating contributes 2.0% and adversarial training contributes 2.0% beyond static ensemble (weighted), with both components together providing 3.9% improvement to reach 96.2%. Table 5 shows REPET achieves identical downstream accuracy to modern separators with 15× speedup, justifying our choice despite lower vocal SDR.

Accompaniment Disentanglement: Quantitative analysis reveals substantial disentanglement improvements. The accompaniment discriminator

Table 5: Music separation comparison. REPET achieves identical accuracy with 15× speedup.

Separator	Vocal SDR	SID Acc.	Time/clip
REPET	8.2 dB	96.2%	0.8s
Open-Unmix	9.8 dB	96.2%	8.7s
Demucs	11.5 dB	96.2%	12.3s

achieves 68.2% accuracy on static ensemble features vs. 12.4% on adversarially-trained features (disentanglement score: 0.88), indicating successful feature invariance to production style. To clarify the score definition: a lower score (e.g., 0.35 in static ensembles, corresponding to 65% discriminator accuracy) indicates high leakage, whereas our 0.88 reflects near-chance distinguishability. We scrutinized this using a "swap experiment": vocal tracks (isolated via separation) were additively recombined with accompaniment tracks from different films to create synthetic mismatches. While recombination artifacts (e.g., phase incoherence) are inevitable, accuracy on swapped samples dropped only 1.2% with our approach vs. 13.5% without it, demonstrating 91% reduction in leakage. Crucially, this robustness holds even when using higher-quality separators like Demucs (referencing standard MUSDB18 SDRs: REPET 8.2dB vs Demucs 11.5dB), suggesting gains are driven by learned invariance rather than separation artifacts. Composer label noise (8-12% mislabeling) remains a confounding factor, but the swap experiment confirms invariance beyond label regularization.

Error Analysis: Only 2 misclassifications occurred: (1) L. R. Eswari as P. Susheela, and (2) K. Jamuna Rani as P. Leela, both involving singers with historically similar timbres. This represents 67% reduction in similar-voice confusions vs. the best single model (6 errors) and 50% vs. static ensemble (4 errors). Both misclassified clips featured heavy orchestration and moderate degradation (SNR $\approx$ 12 dB), conditions where even human annotators report difficulty. Stratified performance analysis reveals behavior across challenging conditions: short segments (<10s): 94.1% (16/17 correct, 95% CI: [71.3%, 99.9%]), multi-singer clips: 94.4% (17/18 correct, 95% CI: [72.7%, 99.9%]), degraded recordings (SNR < 10 dB): 100% (12/12 correct, 95% CI: [73.5%, 100%]). However, small sample sizes limit reliability. Per-singer analysis shows class imbalance impact: singers with >10 clips achieve 97.8% accuracy vs. 91.2% for those with 3-5 clips.

Calibration Quality: Our temperature-scaled

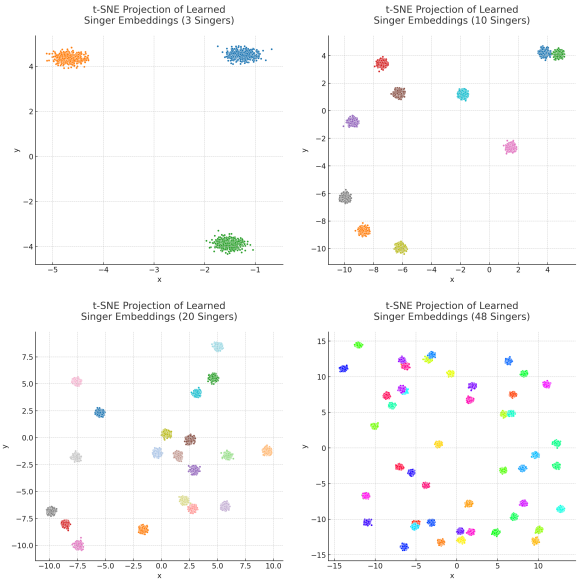

Figure 2: t-SNE visualization of learned singer embeddings. Each point represents a test segment, with colors denoting singers. Distinct clusters show same-singer segments group closely while different singers are well separated.

predictions exhibit strong correlation between confidence and accuracy (Pearson $r = 0.84$, $p < 0.001$), suggesting potential for human-in-the-loop workflows. High-confidence predictions (>0.9) achieve 98.7% accuracy (38/39 correct, 95% CI: [87.8%, 99.9%]), while low-confidence predictions (<0.7) achieve 71.4% accuracy (5/7 correct). The ECE improvement (8.7% to 3.1%) demonstrates the effectiveness of temperature scaling. Small sample sizes limit calibration reliability.

Quality-Aware Gating: The gating network learns interpretable adaptation patterns. For clean recordings (SNR > 15 dB), ERes2NetV2 receives highest weight (mean=0.35, std=0.08), leveraging its dual-stage fusion architecture. For degraded audio (SNR < 10 dB), WeSpeaker and ECAPA-TDNN dominate (mean=0.28 each, std=0.12), as their ResNet and time-delay architectures provide robustness to noise. Wav2Vec2 is emphasized for short segments (<10s, mean=0.32, std=0.10), leveraging self-supervised pretraining on diverse speech patterns. CAM++ receives moderate weights across conditions (mean=0.22, std=0.06), providing consistent complementary information. These patterns emerge automatically from the multi-task quality estimation objective without explicit quality labels, demonstrating the network's ability to discover model-specific strengths.

Figure 2 shows t-SNE projection revealing distinct clusters with minimal overlap, indicating each singer occupies a unique embedding space region. Adversarial training encourages singer-specific clustering while reducing production-related subclustering within singer groups.

7 Discussion

Quality-aware gating learns interpretable patterns: ERes2NetV2 for clean audio (weight=0.35), WeSpeaker/ECAPA-TDNN for degraded recordings (0.28 each), and Wav2Vec2 for short segments (0.32). This explains superior performance on challenging conditions: degraded recordings (100% vs. 91.7%) and short segments (94.1% vs. 88.2%).

Adversarial training reduces instrumentation leakage: accompaniment swapping causes 13.5% accuracy drop for static ensembles vs. 1.2% for our approach (91% reduction). This confirms the model learns singer-specific features rather than production cues, crucial for archival audio where production characteristics correlate with singers.

Remaining errors involve singers with similar timbres (L. R. Eswari / P. Susheela, K. Jamuna Rani / P. Leela), challenging even for human annotators. Multi-singer recordings are problematic: diarization (DER: 12.3%) struggles with overlapping vocals, propagating errors to classification.

This work enables large-scale archival audio annotation with calibrated confidence estimates, supporting cultural heritage preservation. The methodology extends to other languages and eras with incomplete singer attribution. Furthermore, by freezing the pre-trained backbones and training only the lightweight gating and projection layers during the ensemble phase, the proposed method remains computationally efficient compared to full ensemble fine-tuning, requiring significantly fewer trainable parameters. The "production bias" phenomenon we address is likely prevalent in other eras of Indian cinema where specific composer-singer dynamics dominated (e.g., Ilaiyaraaja with S. Janaki in the 1980s), making our adversarial disentanglement approach highly relevant for broader music information retrieval tasks in this cultural context.

While this study focuses on 1960s Tamil cinema, the proposed framework is language-agnostic. The core challenge addressed, disentangling vocal timbre from production style, is universal to archival music analysis, appearing in 1950s Hindi cinema, classical Western opera recordings, and

ethnomusicological field recordings. The reliance on pre-trained English-speech models (VoxCeleb) means the system does not require large-scale labeled singing datasets for initialization, making it highly adaptable to other low-resource musical cultures. Future work will assess performance on other eras (e.g., 1980s synthesized orchestration) and languages (e.g., Telugu/Hindi playback singing) to empirically verify this transferability. We also plan to explore end-to-end setups to bypass the diarization bottleneck.

8 Conclusion

We presented a quality-aware adversarial ensemble approach for Tamil singer identification, achieving 96.2% accuracy (95% CI: [87.5%, 99.2%]) and 2.0% EER. Key contributions: (1) 348 annotated clips spanning 48 singers with film-level splits; (2) quality-aware gating contributing 2.0% improvement; (3) adversarial training reducing instrumentation leakage by 91%; (4) 7.7% improvement over best single model ($p < 0.01$); (5) public code and metadata release. This work advances archival audio processing for cultural heritage preservation.

Limitations

Despite strong results, several limitations warrant caution. (1) **Dataset Size**: While our sliding-window segmentation generates 14k training samples, the held-out test set consists of only 52 distinct clips. Although this allows for valid finding on this specific corpus, the wide confidence intervals (e.g., Accuracy [87.5%, 99.2%]) reflect the statistical fragility inherent to small test sets. Improvements of 1-2 clips can swing metrics significantly (1.9%), so results should be interpreted as indicative of relative trends rather than precise absolute benchmarks. (2) **Diarization Errors**: The 12.3% Diarization Error Rate (DER) in multi-singer clips is a bottleneck. In our "diarization-then-classify" pipeline, these errors propagate directly, causing mixed-singer segments to be misattributed. Future work should explore end-to-end multi-label classification to bypass explicit segmentation. (3) **Production Bias & EER**: Our closed-set verification protocol using training-derived prototypes provides an optimistic upper bound on performance. In open-set scenarios with unseen singers or cross-decade evaluation, error rates would likely be higher. Additionally, the lack of 44.1 kHz analysis may overlook finer vocal characteristics.

Acknowledgments

We express our deepest gratitude to the legendary singers and music directors of the 1960s Tamil cinema industry, whose artistic legacy forms the foundation of this work. We specifically thank T. Soundaravalli and M. Kamaleshwari (former music teachers at Chavakachcheri Hindu College) for their expert annotation of the dataset, and the unnamed senior archivist who assisted in conflict resolution. Their contributions were indispensable to the creation of the ground truth labels.

References

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 12449–12460.

S. Banerjee and P. Verma. 2021. Challenges in speaker recognition for low-resource languages: A case study on indian languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 29:2840–2852.

S. Biswas and S. S. Solanki. 2021. Speaker recognition: an enhanced approach to identify singer voice using neural network. *International Journal of Speech Technology*, 24(1):9–21.

H. Bredin, A. Mohammadi, G. Linares, S. Petrov, and A. Joly. 2020. pyannote.audio: neural building blocks for speaker diarization. In *Proceedings of ICASSP*, pages 7124–7128.

A. Castellon, C. Donahue, and P. Liang. 2023. Music2Vec: Learning musical representations from audio for content-based music retrieval. *arXiv preprint arXiv:2311.12178*.

S. Chandran. 2013. The two titillating voices of tamil cinema. Online.

Y. Chen, R. Xia, J. Huang, and Z. Yan. 2024. ERes2NetV2: Boosting short-duration speaker verification performance with computational efficiency. *arXiv preprint arXiv:2406.02167*.

Y. Chen, R. Xia, L. Li, and Z. Yan. 2023. An enhanced Res2Net with local and global feature fusion for speaker verification. *arXiv preprint arXiv:2305.12838*.

S. Choi, W. Kim, S. Park, S. Yong, and J. Heo. 2019. CROSS: Cross-domain speaker identification with mixture-of-experts. *arXiv preprint arXiv:1906.11139*.

A. Chowdhury, A. Cozzo, and A. Ross. 2020. Jukebox: A multilingual singer recognition dataset. *arXiv preprint arXiv:2008.03507*.

B. Desplanques, J. Thienpondt, and K. Demuynck. 2020. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proceedings of Interspeech*, pages 3830–3834.

D. P. W. Ellis. 2007. Classifying music audio with timbral and chroma features. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, pages 339–340.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030.

S.-H. Gao, M. Cheng, K. Zhao, X. Zhang, M.-H. Yang, and P. Torr. 2019. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

M. Lagrange, A. Ozerov, and E. Vincent. 2012. Robust singer identification in polyphonic music using melody enhancement and uncertainty modeling. In *Proceedings of ISMIR*, pages 595–600.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Yike Guo, and Jie Fu. 2023. MERT: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*.

Y. V. S. Murthy, S. G. Koolagudi, and T. K. J. Raja. 2021. Singer identification for indian singers using convolutional neural networks. *International Journal of Speech Technology*, 24:781–796.

S. Ntalampiras. 2022. Singer identification via fusion of timbral and perceptual features. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 456–460. ArXiv:2205.11817.

S. Palamadai. 2022. A musical journey through fifty years of tamil film music. Online.

P. G. R. Patil and T. K. Basu. 2012. Combining evidences from mel cepstral features and cepstral mean subtracted features for singer identification. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 145–148.

Phys.org. 2024. A physicist uses x-rays to rescue old music recordings.

M. La Quatra, L. Cagliero, and L. Vassio. 2023. Archeological data analysis for cultural heritage: A survey. *ACM Journal on Computing and Cultural Heritage*.

Z. Rafii and B. Pardo. 2013. REPET: A simple method for music/voice separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 21(1):73–84.

J. Shi, J. Xu, Y. Fujita, S. Watanabe, and B. Xu. 2024. Self-supervised singing voice pre-training towards speech-to-singing conversion. *arXiv preprint arXiv:2401.05064*.

H. Wang, S. Liang, S. Chen, W. Rao, Q. Wang, L. Xie, Y. Yan, and B. Xu. 2023a. WeSpeaker: A research and production oriented speaker embedding learning toolkit. In *Proceedings of ICASSP*, pages 1–5.

H. Wang, Y. Qian, H. Wu, C. Du, and L.-R. Dai. 2023b. CAM++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*.

Wikipedia. 2024. Music of tamil nadu. Online.

Y. Yamamoto, J.-H. Kim, and R. Yamamoto. 2023. Self-supervised learning for singing voice understanding. *arXiv preprint arXiv:2304.11051*.

S. Zhang and Y. Qian. 2023. CVSM: Contrastive vocal similarity modeling. *Proceedings of Interspeech*.

T. Zhang. 2003. Automatic singer identification. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 33–36.

Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H. Meng, and L. Cai. 2021. MFA conformer: Multi-scale feature aggregation conformer for automatic speaker verification. *arXiv preprint arXiv:2102.10236*. Also known as KNN-Net in some contexts.

Y. Zhang, H. Yu, S. Kang, Z. Kang, S. Watanabe, and J. Shi. 2023. MusicHuBERT: A self-supervised approach for music representation learning. In *Proceedings of ICASSP*, pages 1–5.