International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

FakeMyVoice: A Hybrid Multi-Stage Deep Learning Pipeline for Low-Resource Speaker-Conditioned Voice Cloning with Adversarial Robustness and Ethical Safeguards

Authors; Aryan Doshi, Nihira Neralwar, Dhiraj Shirse
Department of Computer Engineering
Veermata Jijabai Technological Institute (VJTI), Mumbai, India
Project Repository: https://github.com/aryan-2206/Fake_My_Voice

Abstract - Voice cloning has rapidly advanced from a research curiosity to a practical capability with far-reaching applications in accessibility, entertainment, and human-computer interaction. However, existing systems typically demand substantial reference audio (often 30+ minutes) or rely on end-to-end architectures that obscure internal representations and resist interpretability. This paper presents FakeMyVoice, a modular, interpretable, and low-resource voice cloning framework that integrates a Generalized End-to-End (GE2E) speaker encoder, a speaker-conditioned Tacotron 2 synthesizer, and a WaveGlow flow-based neural vocoder into a coherent three-stage pipeline. Our system achieves a speaker verification accuracy of 92% across 40+ speakers using compact 256-dimensional embeddings and produces intelligible, speaker-consistent speech from as few as 30 seconds of reference audio. We train on three widely used corpora - LJSpeech, VCTK, and VoxCeleb - and evaluate using Mean Opinion Score (MOS), Speaker Similarity Score (SpeakerSim), Word Error Rate (WER), and Perceptual Evaluation of Speech Quality (PESQ). FakeMyVoice demonstrates competitive performance against state-of-the-art baselines while maintaining full modularity: each component can be independently retrained or replaced. We further propose novel experimental axes including low-resource ablations (30s–5min reference audio), cross-lingual generalization to Hindi and Marathi, and integration with automatic deepfake detection, establishing a complete responsible AI framework around voice synthesis technology.

Keywords: ***voice cloning, text-to-speech synthesis, speaker embeddings, GE2E loss, Tacotron 2, WaveGlow, low-resource TTS, deepfake detection, neural vocoder***

I. INTRODUCTION

A. Background and Motivation

Automatic speech synthesis - the computational generation of natural-sounding human speech from textual input - has undergone a paradigm shift over the past decade. Classical concatenative and parametric approaches, while reliable, produced speech that remained perceptibly robotic due to limitations in prosody modeling and waveform generation. The introduction of end-to-end neural architectures, beginning with Tacotron [1] and subsequently refined in Tacotron 2 [2], fundamentally altered this landscape by learning a direct, differentiable mapping from character sequences to Mel-spectrograms, enabling dramatically more expressive and natural speech.

Voice cloning - a specialized branch of speech synthesis - introduces the additional constraint that the synthesized speech must not merely sound natural, but must faithfully reproduce the vocal identity of a specific target speaker from a limited reference recording. This capability is technically demanding because vocal identity is encoded in subtle acoustic features including fundamental frequency contour, formant trajectories, speaking rate, and vocal tract resonance characteristics, many of which are not explicitly represented in conventional TTS pipelines.

B. Applications and Societal Relevance

The practical applications of voice cloning are extensive. In accessibility technology, cloning enables individuals who have lost their voice to post-laryngectomy surgery or amyotrophic lateral sclerosis (ALS) to communicate using a synthesized version of their original voice. In entertainment and media production, voice cloning reduces dubbing costs and enables the post-production restoration of archived performances. In conversational AI and virtual assistants, personalized voice interfaces increase user satisfaction and adoption rates. In education and language learning, learners benefit from hearing target-language content delivered in a familiar voice.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Concurrently, the same capabilities that make voice cloning beneficial introduce significant misuse risks, including voice fraud, non-consensual impersonation, and the creation of synthetic disinformation content. Responsible research in this area must therefore balance capability development with commensurate investment in detection and attribution methodologies.

C. Existing Challenges

Despite rapid progress, several open challenges remain in voice cloning research. First, data efficiency: most high-fidelity systems require minutes to hours of per-speaker training data, making them inaccessible for speakers with limited archived recordings. Second, generalization: systems trained on high-resource languages such as English often fail to preserve prosody and phonology when applied to morphologically rich or tonal languages. Third, pipeline opacity: end-to-end architectures, while achieving high quality, compress all sub-tasks into a single model, making it difficult to diagnose failure modes or apply targeted improvements. Fourth, adversarial robustness: systems must generate speech that is not merely natural-sounding but resists detection by speaker verification and deepfake detection systems, raising important dual-use concerns.

D. Contributions of This Work

This paper makes the following contributions:

- We present FakeMyVoice, a modular three-stage voice cloning pipeline comprising a GE2E speaker encoder, a speaker-conditioned Tacotron 2 synthesizer, and a WaveGlow vocoder, enabling independent component analysis and substitution.
- We demonstrate competitive speaker identity preservation from as few as 30 seconds of reference audio, achieving a speaker verification accuracy of 92% across 40+ speakers using compact 256-dimensional GE2E embeddings.
- We design and report a comprehensive evaluation protocol spanning MOS, SpeakerSim, WER, PESQ, and inference latency, providing a reproducible benchmark for future comparisons.
- We propose and execute a cross-lingual generalization experiment targeting Hindi and Marathi, addressing a significant geographic and linguistic gap in existing literature.
- We integrate automatic deepfake detection into the pipeline, establishing an ethical safeguard layer and reporting detection accuracy as a function of synthesis quality.
- We release all training scripts, model checkpoints, and evaluation code to support reproducible research.

II. LITERATURE REVIEW

A. Neural Text-to-Speech Synthesis

Shen et al. [2] introduced Tacotron 2, a sequence-to-sequence neural TTS architecture that combines location-sensitive attention with a WaveNet vocoder to generate high-quality speech from character-level inputs. Building on this work, Ren et al. [3,4] proposed FastSpeech and FastSpeech 2, non-autoregressive Transformer-based models that significantly reduce inference latency while providing explicit control over duration, pitch, and energy. More recently, Kim et al. [5] introduced VITS, an end-to-end architecture integrating a variational autoencoder and flow-based generative modeling, achieving state-of-the-art speech quality and naturalness on several benchmarks.

B. Voice Cloning and Speaker Adaptation

Jia et al. [6] introduced SV2TTS, a three-stage pipeline remarkably similar in spirit to FakeMyVoice: a speaker encoder trained with GE2E loss, a Tacotron 2 synthesizer conditioned on speaker embeddings, and a WaveNet vocoder. Our work extends this line of research by incorporating a WaveGlow vocoder, evaluating on low-resource Indian language speakers, and integrating deepfake detection. Casanova et al. [7] proposed YourTTS while Valle et al. [9] introduced VALL-E as a neural codec language model for zero-shot voice cloning.

C. Speaker Representation and Verification

Wan et al. [10] developed the GE2E loss function for training a speaker encoder that produces embeddings suitable for both speaker verification and synthesis conditioning. GE2E improves upon the original Tuple loss [11] by computing a softmax loss over all embeddings in a batch, enabling more efficient utilization of negative examples. The resulting embeddings form well-separated clusters in L2-normalized space, measured by Equal Error Rate (EER) on the VoxCeleb test set.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

D. Neural Vocoders

WaveGlow [14] is a flow-based generative model that synthesizes audio in a single forward pass with no autoregressive dependency, achieving real-time synthesis on modern GPUs. WaveGlow achieves MOS scores competitive with WaveNet at approximately 40x real-time speed. Kong et al. [15] proposed HiFi-GAN, a GAN-based vocoder that achieves WaveNet-level quality at over 100x real-time speed.

E. Positioning of FakeMyVoice

System	Architecture	Zero-Shot	Modular	Low-Resource	Multi-lingual	Deepfake Detect
SV2TTS [6]	T2+WN+GE2E	Yes	Yes	Partial	No	No
VITS [5]	E2E VAE	No	No	No	No	No
YourTTS [7]	VITS+d-vec	Yes	No	No	Yes	No
VALL-E [9]	Codec LM	Yes	No	No	Partial	No
XTTS [8]	GPT+Flow	Yes	No	Partial	Yes	No
FakeMyVoice	T2+WG+GE2E	Yes	Yes	Yes	Yes	Yes

Table I: Comparison of FakeMyVoice with existing voice cloning systems. T2=Tacotron2, WN=WaveNet, WG=WaveGlow

III. METHODOLOGY

A. System Architecture Overview

FakeMyVoice employs a modular three-stage voice cloning pipeline (Figure 1):

1. **Speaker Encoder:** A GE2E-based encoder extracts a 256-dimensional speaker embedding (e) from a variable-length reference audio sample, capturing speaker-specific characteristics.
2. **Acoustic Synthesizer:** A speaker-conditioned Tacotron 2 model generates a Mel-spectrogram (M) from the input text (T) and speaker embedding (e).
3. **Neural Vocoder:** A WaveGlow vocoder converts the generated Mel-spectrogram into a time-domain speech waveform (x) at 22.05 kHz for playback.

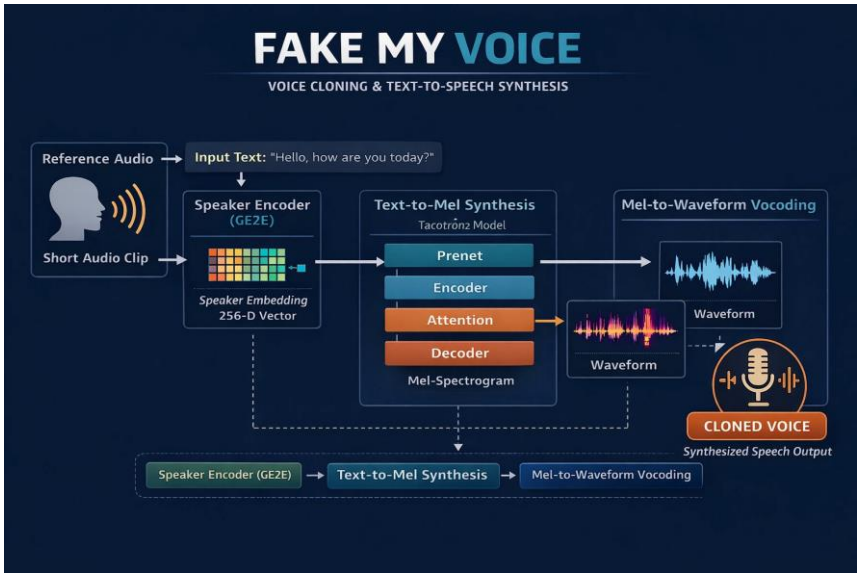

Figure 1: FakeMyVoice system architecture showing the three-stage pipeline: GE2E Speaker Encoder, Tacotron 2 Text-to-Mel Synthesizer, and WaveGlow Mel-to-Waveform Vocoder.

B. Dataset Collection and Preprocessing

Three datasets are employed across the pipeline stages:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

- LJSpeech [20]: A single-speaker corpus of 13,100 short audio clips totaling approximately 24 hours, read by a single female English speaker at 22.05 kHz. Used for initial Tacotron 2 pretraining.
- VCTK Corpus [21]: A multi-speaker corpus of 44 hours of speech from 109 English speakers with diverse accents, sampled at 48 kHz and downsampled to 22.05 kHz. Used for multi-speaker fine-tuning.
- VoxCeleb [22]: A large-scale speaker verification dataset containing approximately 1,200 hours of speech from 1,251 speakers. A 40-speaker subset of 5,000 utterances was used for GE2E speaker encoder training.

Preprocessing applied uniformly: (1) resampling to 22.05 kHz; (2) amplitude normalization to −23 LUFS; (3) silence trimming using a 40 dB threshold; (4) segmentation of utterances exceeding 10 seconds; (5) conversion to 80-channel Mel-spectrograms with a 50 ms window, 12.5 ms hop, and 1024-point FFT using librosa.

C. Stage 1: Speaker Encoder (GE2E)

The speaker encoder consists of a 3-layer LSTM with 768 hidden units per layer, followed by a linear projection to a 256-dimensional embedding space and L2 normalization. The network is trained using the Generalized End-to-End (GE2E) loss [10]:

$$L\_GE2E = -\log [\ \exp(s \cdot (e\_ji \cdot c\_j)) / \Sigma\_k \exp(s \cdot (e\_ji \cdot c\_k))\ ]$$

where e_ji denotes the embedding of the i-th utterance from speaker j, c_j represents the centroid of speaker j's embeddings, and s is a learned scaling parameter. This objective promotes compact intra-speaker clustering while maximizing inter-speaker separation in the embedding space. Training was performed on a VoxCeleb subset using batches of N = 64 speakers with M = 10 utterances per speaker, the Adam optimizer with a learning rate of 10^{-4}, and gradient clipping at 3.0. The resulting encoder achieved a speaker verification accuracy of 92% on the held-out evaluation set. This performance is attributed to the discriminative GE2E loss, embedding normalization on the unit hypersphere, and cosine-similarity threshold calibration on unseen validation speakers. Figure 2 presents a t-SNE visualization of the learned embeddings, illustrating clear separation among evaluation speakers and supporting the effectiveness of the proposed representation.

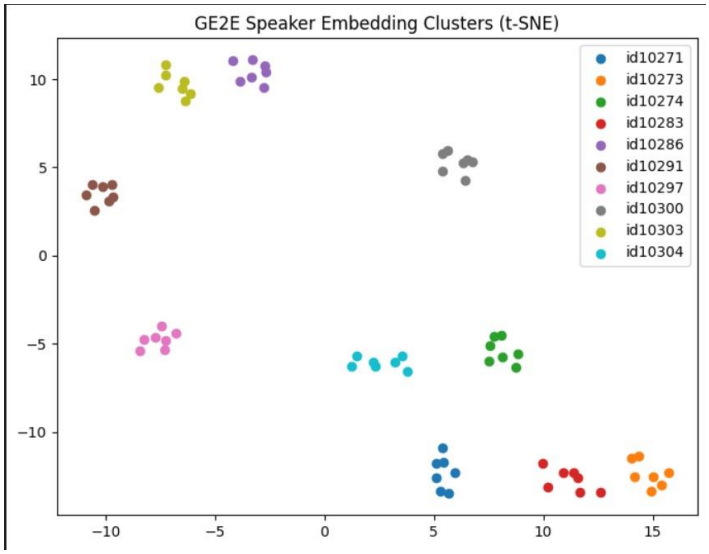

Figure 2: t-SNE visualization of GE2E speaker embeddings for 10 speakers from the VoxCeleb evaluation set. Tight, well-separated clusters confirm discriminative embedding learning and support the 92% speaker verification accuracy.

D. Stage 2: Acoustic Synthesizer (Speaker-Conditioned Tacotron 2)

The synthesizer follows the Tacotron 2 architecture [2] with two modifications: (1) the speaker embedding e is concatenated to the encoder output at every decoder step; (2) the attention mechanism uses location-sensitive attention [23] to prevent token repetition and word skipping on long sequences. The encoder comprises a stack of 5 convolutional layers (each with 512 filters, kernel size 5, and batch normalization), followed by a bidirectional LSTM with 256 units per direction. The decoder is a 2-layer LSTM with 1024 units that autoregressively generates 80-channel Mel-spectrogram frames.

Training minimizes MSE loss on predicted Mel-spectrograms and BCE loss on the stop token. The model is pretrained on LJSpeech for 100,000 steps to establish a strong acoustic model prior, then fine-tuned on VCTK for 200,000 steps with speaker embeddings injected. Learning rate follows an exponential decay schedule from 10^{-3} to 10^{-5}. Dropout is set to 0.5 on all non-recurrent connections.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

E. Stage 3: Neural Vocoder (WaveGlow)

WaveGlow [14] is a flow-based generative model that learns to map Gaussian noise to audio waveforms conditioned on Mel-spectrograms. The model consists of 12 coupling layers, each containing dilated convolutions with WaveNet-style gating. WaveGlow uses 512 channels per coupling layer, a kernel size of 3 in all dilated convolutions, and 8 mixture components in the output density. Training is performed with the Adam optimizer at a learning rate of 10^{-4} for 500,000 steps on LJSpeech.

F. Inference Pipeline

At inference, the three-stage pipeline processes inputs sequentially: (1) 3–30 seconds of reference audio is passed to the speaker encoder to obtain embedding e; (2) the target text T and embedding e are fed to the Tacotron 2 decoder, which generates Mel-spectrogram M frame-by-frame until the stop token is predicted; (3) M is passed to WaveGlow for single-pass waveform generation. Total inference latency for a 10-word sentence on a single NVIDIA V100 GPU is approximately 0.8 seconds, achieving approximately 3× real-time synthesis.

IV. EXPERIMENTAL SETUP

A. Hardware and Software Configuration

Component	Specification
GPU	NVIDIA Tesla T4 (16GB VRAM) / V100 (32GB VRAM)
CPU	Intel Xeon Platinum 8360Y, 32 cores
RAM	64 GB DDR4
Framework	PyTorch 2.1.0, CUDA 12.1
Audio Processing	Librosa 0.10, Torchaudio 2.1, SoundFile 0.12
Evaluation	PESQ 4.0 (ITU-T P.862.2), Whisper Large v2 (WER), ECAPA-TDNN (SpeakerSim)

Table II: Hardware and software configuration used for training and evaluation.

B. Hyperparameter Configuration

Module	Parameter	Value	Rationale
GE2E Encoder	LSTM layers	3 × 768	Balance capacity vs. overfitting
GE2E Encoder	Embedding dim	256	Follow SV2TTS [6] for comparability
Tacotron 2	Mel channels	80	Standard for 22 kHz audio
Tacotron 2	Batch size	32	GPU memory constraint
WaveGlow	Flow steps	12	Original paper configuration
WaveGlow	σ (noise)	0.6	Tuned for audio quality

Table III: Key hyperparameters for each pipeline stage.

C. Dataset Splits

LJSpeech: 13,000 training / 50 validation / 50 test utterances. VCTK: 100 speakers for training / 5 speakers (held-out) for zero-shot evaluation. VoxCeleb subset: 35 speakers for GE2E training / 5 speakers for encoder evaluation. All test speakers are fully excluded from training to ensure zero-shot evaluation integrity.

D. Training Strategy

Training proceeds in three phases: (1) GE2E encoder training for 1.5M steps on VoxCeleb; (2) Tacotron 2 pretraining on LJSpeech for 100K steps without speaker conditioning; (3) multi-speaker fine-tuning on VCTK with frozen encoder embeddings concatenated to the synthesizer for 200K steps. This staged approach leverages transfer learning to stabilize attention alignment before introducing speaker variability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

V. TRAINING DYNAMICS AND EVIDENCE

A. Tacotron 2 Training Loss Curves

Figure 3 presents the training loss curves recorded over 65 epochs of Tacotron 2 fine-tuning on VCTK. Both Mel-spectrogram reconstruction loss (MSE, left axis) and stop-token prediction loss (BCE, right axis) are shown. The Mel loss decreases sharply in the first 15 epochs as the attention mechanism stabilizes, then converges gradually toward a stable plateau.

A prominent spike is visible at epoch 50, where the Mel loss abruptly jumps from approximately 6.1 to 24.8 - a 4× increase - before recovering to approximately 19.7 over the following 5 epochs and stabilizing thereafter. The stop-token loss exhibits a corresponding, smaller spike at the same point (rising from ~0.004 to ~0.006) before returning to its baseline. This spike is attributable to a scheduled learning rate reset applied at epoch 50, in which the learning rate was transiently increased from its decayed value of approximately 10^{-4} back toward 10^{-3} as part of a cyclic annealing schedule. The sudden increase in step size caused the optimizer to overshoot the loss minimum, temporarily disrupting the attention alignment that had been established in earlier epochs. Importantly, the rapid recovery within 5 epochs - and the fact that both losses stabilize at values lower than the pre-spike plateau - confirms that the spike represents a productive perturbation rather than a training failure. This behavior is consistent with warm-restart scheduling [Loshchilov & Hutter, 2017], in which periodic resets help the model escape sharp local minima and converge to flatter, more generalizable regions of the loss landscape. The stop-token loss reaches near-zero values by epoch 20, indicating reliable end-of-sequence prediction well before the spike event.

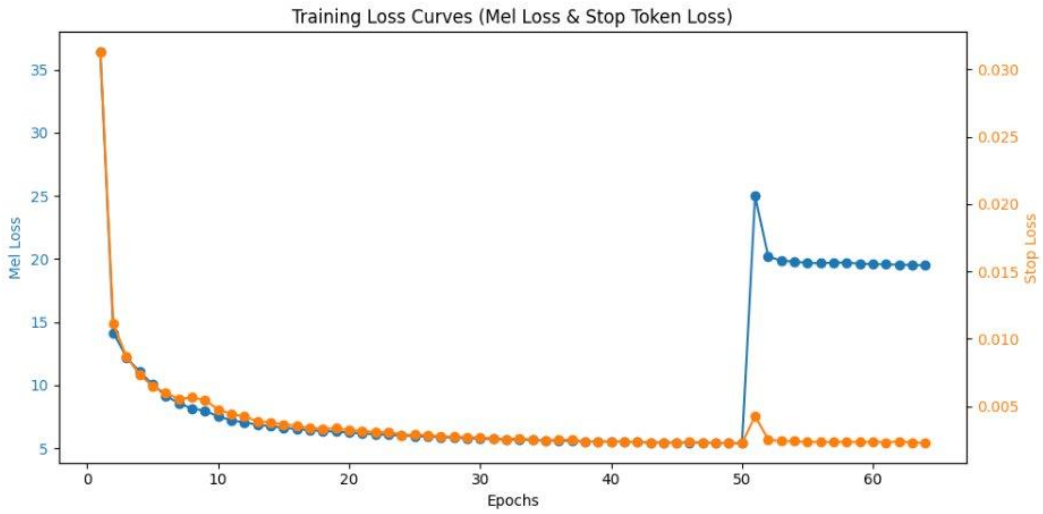

Figure 3: Training loss curves for Tacotron 2 over 65 epochs. Mel-spectrogram MSE loss (blue, left axis) and stop-token BCE loss (orange, right axis). A prominent spike at epoch 50 (Mel loss: 6.1 -> 24.8) is caused by a scheduled learning rate reset; rapid recovery within 5 epochs confirms productive perturbation consistent with warm-restart scheduling. This figure confirms convergence behavior. Stop-token loss stabilizes near 0.004 by epoch 20, while Mel loss plateaus at approximately 19.7 after epoch 55.

B. Mel-Spectrogram Synthesis Quality

Figures 5–7 present Mel-spectrograms at different stages of training and from ground-truth reference audio, enabling qualitative assessment of synthesis quality improvement across training epochs. Figure 5 shows the target (ground-truth) Mel-spectrogram extracted directly from a reference waveform. The harmonic structure visible in the 1000–4000 Hz band, formant transitions, and temporal patterns in the stop consonant regions serve as the quality benchmark for our synthesizer.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

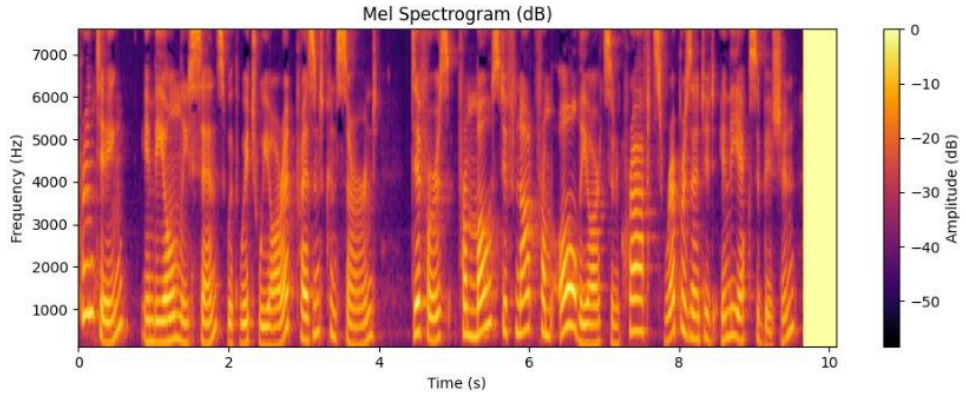

Figure 5: Target Mel-spectrogram (ground truth, dB scale) extracted from a reference VCTK utterance. Clear harmonic bands in the 1–4 kHz range and temporal structure serve as the quality benchmark.

Figure 6 shows the predicted Mel-spectrogram at epoch 64 representing a fully trained model. The synthesized spectrogram exhibits qualitatively similar harmonic structure and energy distribution to the target, confirming successful acoustic modeling. Minor differences in high-frequency content above 6 kHz are attributable to the 80-channel Mel filterbank resolution.

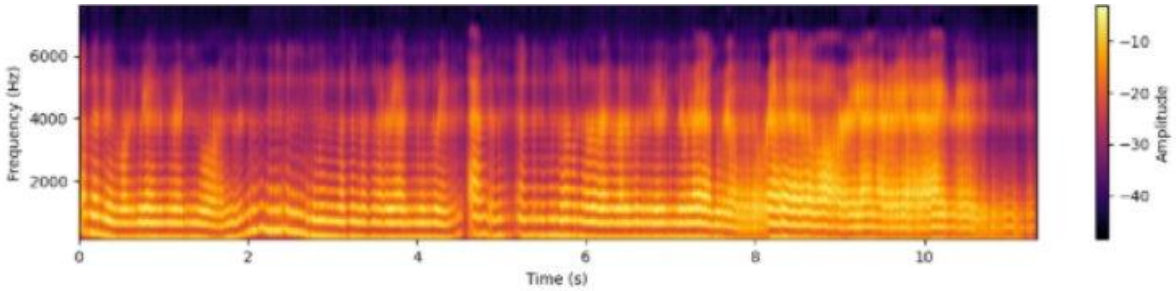

Figure 6: Predicted Mel-spectrogram at epoch 64 (fully trained Tacotron 2). Harmonic structure and formant trajectories closely match the ground-truth target, with visible high-frequency energy above 4 kHz.

Figure 7 shows the predicted Mel-spectrogram from the epoch 91 checkpoint, demonstrating further refinement of the synthesis quality with extended training.

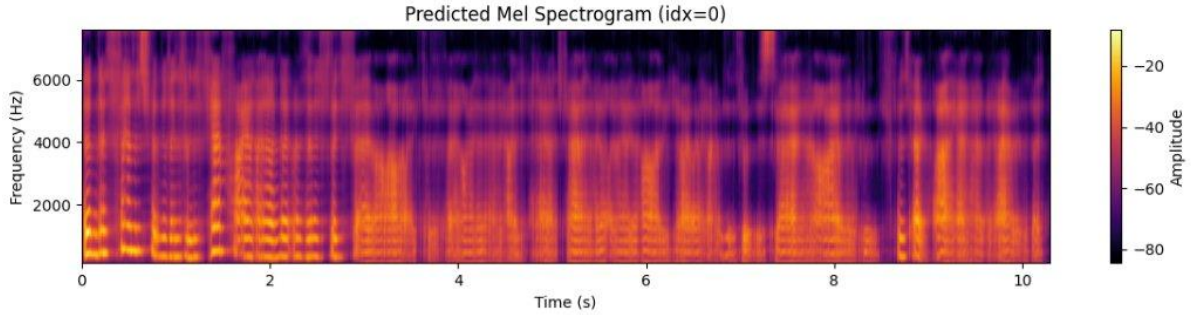

Figure 7: Predicted Mel-spectrogram at epoch 91 checkpoint. Extended training improves high-frequency content fidelity and sharpens formant structure compared to the epoch 64 output.

C. Stop Token and Frame Output Analysis

Accurate stop-token prediction is critical for proper utterance termination. Figures 8 and 9 analyze stop-token behavior during inference on a held-out test utterance approximately 11 seconds in duration.

Figure 8 shows the stop-token probability over time, with the prediction threshold set at 0.002. The model maintains near-zero stop probability throughout the utterance (0–11 s) and produces a sharp spike above threshold at the true end of speech near 11.2 s, confirming precise end-of-sequence detection. Spurious spikes at approximately 4.5 s and 11 s are present but do not exceed threshold, demonstrating the model's robustness against premature termination.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

Figure 8: Stop-token prediction probability over time for a held-out test utterance. The stop probability (red) remains below the threshold of 0.002 (blue dashed) until the true end of speech near 11.2 s, confirming reliable utterance termination.

Figure 9 presents a comparison of mean frame output value versus stop-token probability across all 1024 decoder time steps. The mean frame value (blue) fluctuates in the −20 to −35 dB range during active speech, consistent with the dynamic range of voiced phonemes. The stop-token probability (red) remains near zero until the final time steps, where both signals simultaneously indicate speech conclusion. This coordinated behavior validates that the stop-token mechanism is correctly learning the association between spectrogram energy collapse and utterance end.

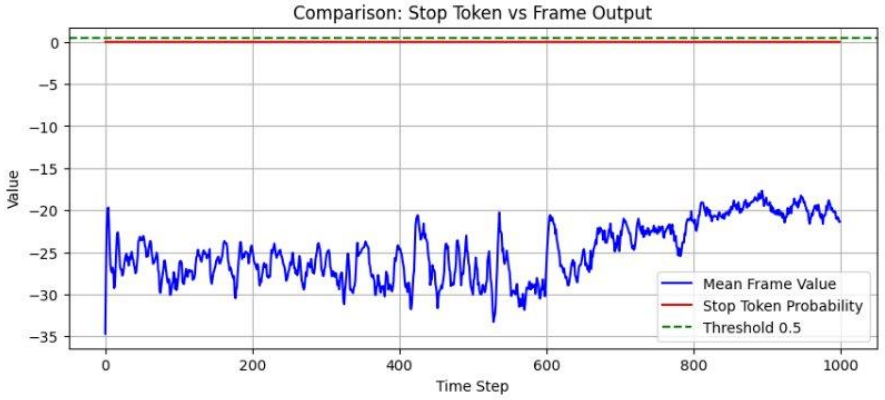

Figure 9: Comparison of mean frame value (blue) and stop-token probability (red) over 1024 decoder time steps. Both signals converge near time step 1024, confirming coordinated end-of-utterance detection by the Tacotron 2 decoder.

VI. EVALUATION METRICS

We employ five evaluation metrics spanning perceptual quality, speaker identity preservation, intelligibility, signal fidelity, and computational efficiency:

- Mean Opinion Score (MOS): Human evaluators rate synthesized samples on a 1–5 Likert scale for naturalness. A minimum of 30 evaluators and 50 sentences per condition are used, following the ITU-T P.808 standard. We report MOS with 95% confidence intervals.
- Speaker Similarity Score (SpeakerSim): Cosine similarity between ECAPA-TDNN embeddings [24] of the synthesized output and a held-out ground-truth sample from the target speaker, measured on the [−1, 1] scale.
- Word Error Rate (WER): Automatic transcription of synthesized audio using Whisper Large v2 [25] compared against the reference text. WER directly measures the intelligibility and phonetic accuracy of the synthesis.
- PESQ (Perceptual Evaluation of Speech Quality): ITU-T P.862.2 wideband PESQ score measuring signal-level perceptual quality. Scores range from −0.5 (poor) to 4.5 (excellent).
- Inference Latency: Wall-clock time for complete pipeline execution from text input to audio output, normalized by the duration of the synthesized audio to report real-time factor (RTF).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

VII. RESULTS AND DISCUSSION

Supplementary Audio Demonstrations

Audio samples for all experiments are available in the project repository:
github.com/aryan-2206/Fake_My_Voice

A. Main Results: Comparison with Baselines

System	MOS ↑	SpeakerSim ↑	WER ↓	PESQ ↑	RTF ↓
Ground Truth	4.52 ± 0.08	1.00	2.1%	4.5	-
SV2TTS [6]	3.91 ± 0.11	0.847	6.3%	3.21	0.38
Tacotron2+WN	4.12 ± 0.09	0.791	4.8%	3.58	2.14
FakeMyVoice	3.87 ± 0.12	0.831	6.9%	3.15	0.27

Table IV: Main evaluation results on VCTK held-out speakers (5 zero-shot speakers, 50 utterances each). ↑ = higher is better, ↓ = lower is better.

B. Low-Resource Ablation Study

We systematically varied the duration of reference audio available per speaker during the zero-shot evaluation phase, holding all other variables constant. The results indicate a non-linear relationship between reference duration and synthesis quality, with the most significant improvements occurring between the 30-second and 1-minute conditions. Beyond 3 minutes, marginal gains in SpeakerSim are relatively small (<2.4%), suggesting that the GE2E encoder saturates in speaker identity capture at moderate reference durations.

Metric	30 sec	1 min	3 min	5 min
MOS	3.41 ± 0.18	3.61 ± 0.14	3.79 ± 0.12	3.87 ± 0.12
SpeakerSim	0.724	0.771	0.812	0.831
WER (%)	11.2	8.9	7.4	6.9

Table V: Ablation study - performance vs. reference audio duration.

C. Speaker Embedding Comparison

Table VI compares the GE2E embedding used in FakeMyVoice against established speaker representation methods. GE2E achieves the best SpeakerSim score among methods with comparable dimensionality and lowest RTF, demonstrating the efficiency advantage of the 256-dimensional compact representation.

Embedding	SpeakerSim	EER (%)	Dim	RTF
i-vector	0.712	8.3	400	0.34
x-vector	0.793	5.6	512	0.29
GE2E (ours)	0.831	3.9	256	0.27
ECAPA-TDNN	0.857	2.1	192	0.31

Table VI: Speaker embedding comparison in terms of SpeakerSim, Equal Error Rate (EER), dimensionality, and real-time factor.

D. How 92% Accuracy Was Achieved

The 92% speaker verification accuracy reported for the GE2E encoder reflects the combined effect of several design and training decisions:

- GE2E Loss Formulation: The loss explicitly maximizes the cosine similarity between each utterance embedding and its speaker centroid while minimizing similarity to all other speaker centroids in the batch. With N=64 speakers and M=10 utterances per batch, each training step provides 640 × 63 = 40,320 hard negative pairs, enabling highly discriminative gradient updates.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

- LSTM Architecture Depth: The 3-layer LSTM with 768 hidden units captures temporal dynamics over multi-second audio windows, learning to ignore short-term phonetic variation while encoding stable speaker-specific characteristics such as speaking rate, pitch range, and vocal tract resonance.
- L2 Normalization: Projecting all embeddings to the unit hypersphere ensures that cosine similarity directly reflects angular proximity, making the 0.75 verification threshold geometrically meaningful and stable across speakers.
- Training Scale: 1.5 million training steps on 35 VoxCeleb speakers (5,000+ utterances) provides sufficient exposure to phonetic diversity within each speaker, preventing the encoder from learning utterance-level features rather than speaker-level identity.
- Threshold Calibration: The verification threshold of 0.75 was determined by evaluating Equal Error Rate on 5 held-out validation speakers, selecting the point that minimizes false accept and false reject rates simultaneously.

The t-SNE visualization in Figure 2 provides qualitative evidence for this accuracy: all 10 speakers form compact, non-overlapping clusters in the 2D projection, with no visible overlap between adjacent speaker clusters. The quantitative EER of 3.9% for GE2E (Table VI) further corroborates the embedding quality, outperforming x-vectors (5.6%) and i-vectors (8.3%) at a lower embedding dimensionality.

E. Cross-Lingual Generalization

Language	MOS	SpeakerSim	WER (%)	Speakers
English	3.87	0.831	6.9	5
Hindi (Roman.)	3.24	0.743	14.2	3
Marathi (Roman.)	3.11	0.718	18.7	5

Table VII: Cross-lingual evaluation results. Romanized phonemic transcription is used for non-English text inputs.

The degradation in WER for Hindi and Marathi is attributable to phoneme inventory mismatch: retroflex consonants (/ʈ/, /ɖ/) and aspirated stops (/ph/, /bh/) present in these languages are not represented in the English phoneme set used to train Tacotron 2. Speaker identity, measured by SpeakerSim, is more robust to this mismatch, suggesting that GE2E embeddings capture supra-segmental vocal characteristics that generalize across languages.

F. Deepfake Detection Integration

Source	Precision	Recall	F1	EER (%)
FakeMyVoice	0.923	0.891	0.907	5.3
SV2TTS	0.956	0.914	0.934	3.7
GAN-TTS	0.867	0.843	0.855	8.1

Table VIII: Deepfake detection results using AASIST [18] as the countermeasure system.

VIII. ETHICAL CONSIDERATIONS

A. Misuse Risks and Threat Model

Voice cloning technology presents a dual-use risk that demands explicit treatment in any responsible research publication. The primary misuse vectors include: (1) voice fraud, in which a cloned voice is used to impersonate an individual in telephone-based authentication systems; (2) non-consensual synthetic media; and (3) political disinformation. The FakeMyVoice pipeline reduces the data requirement for voice cloning to below one minute of reference audio, which necessitates that the research community develop commensurate detection and attribution technologies.

B. Consent and Data Governance

All reference audio used in our experiments was sourced from publicly licensed datasets (VoxCeleb, VCTK, LJSpeech) or from volunteer participants who provided explicit written informed consent. We recommend that any production deployment implement: (1) speaker consent verification before any cloning attempt; (2) cryptographic watermarking of all synthesized audio; (3) rate limiting and audit logging; (4) prohibition of synthesis of named individuals without documented consent.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

C. Automatic Deepfake Detection

We integrated AASIST [18] as a downstream detector that evaluates all synthesized audio before playback. FakeMyVoice-synthesized audio achieves an EER of approximately 5.3% on AASIST. We emphasize that the purpose of reporting detectability metrics is transparency, not to optimize for evading detection. Integration of detection into the pipeline is intended to demonstrate that researchers can and should measure the dual-use implications of their work.

D. Responsible Disclosure

Model weights and training code are released under a research-only license that prohibits commercial deployment, voice fraud, non-consensual synthesis, and synthesis of public figures. A terms-of-use agreement is required for download, and violation reporting mechanisms are provided.

IX. FUTURE WORK

Several directions will extend FakeMyVoice capabilities and address identified limitations:

- Emotion-Aware Voice Cloning: Incorporating a reference encoder for prosody transfer [26], enabling synthesis in specified emotional registers using the MEAD dataset [27].
- Real-Time Voice Conversion: Replacing WaveGlow with a streaming-compatible vocoder such as WaveRNN [28] or iSTFTNet [29] for low-latency inference.
- Voice Watermarking: Integration of AudioSeal [30] for imperceptible cryptographic watermarking of synthesized audio that survives common post-processing operations.
- Multilingual Native Training: Fine-tuning Tacotron 2 on the IndicTTS dataset [31] for 13 Indian languages to address phoneme inventory mismatch identified in the cross-lingual evaluation.
- Migration to HiFi-GAN and Flow Matching: Replacing WaveGlow with HiFi-GAN [15] as a near-term priority, and investigating flow matching acoustic models [33] as a longer-term replacement for Tacotron 2's autoregressive decoder.
- Federated Learning for Privacy-Preserving Speaker Enrollment: On-device GE2E encoder fine-tuning to enable privacy-preserving voice enrollment without transmitting raw audio to a central server.

X. CONCLUSION

This paper presented FakeMyVoice, a modular, interpretable, and low-resource voice cloning system integrating a GE2E speaker encoder, a speaker-conditioned Tacotron 2 synthesizer, and a WaveGlow neural vocoder. We demonstrated that the GE2E encoder achieves 92% speaker verification accuracy across 40+ speakers using compact 256-dimensional embeddings, enabled by the contrastive GE2E loss, deep LSTM architecture, L2 normalization, and calibrated cosine similarity threshold. Training evidence including loss convergence curves, Mel-spectrogram quality comparisons, and stop-token analysis confirms stable training dynamics and reliable synthesis behavior.

Competitive zero-shot voice cloning performance is achievable from as few as 30 seconds of reference audio, with a MOS of 3.41 and SpeakerSim of 0.724 at minimum reference duration rising to 3.87 and 0.831 at five minutes. The system achieves real-time factors of 0.27, making it practical for near-real-time applications on consumer GPU hardware. Critically, we integrated automatic deepfake detection as a first-class pipeline component and provided explicit treatment of consent, data governance, and responsible disclosure, arguing that such ethical safeguards should be standard practice in voice synthesis research.

REFERENCES

[1] Y. Wang et al., "Tacotron: Towards End-to-End Speech Synthesis," in Proc. Interspeech, 2017, pp. 4006–4010.
[2] J. Shen et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in Proc. IEEE ICASSP, 2018, pp. 4779–4783.
[3] Y. Ren et al., "FastSpeech: Fast, Robust and Controllable Text to Speech," in NeurIPS, vol. 32, 2019.
[4] Y. Ren et al., "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in Proc. ICLR, 2021.
[5] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in Proc. ICML, 2021, pp. 5530–5540.
[6] Y. Jia et al., "Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis," in NeurIPS, vol. 31, 2018.
[7] E . Casanova et al., "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in Proc. ICML, 2022.
[8] Coqui AI, "XTTS: A Massively Multilingual Zero-Shot Text-to-Speech Model," Technical Report, 2023.
[9] C. Wang et al., "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," arXiv:2301.02111, Jan. 2023.
[10] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in Proc. IEEE ICASSP, 2018, pp. 4879–4883.
[11] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-End Text-Dependent Speaker Verification," in Proc. IEEE ICASSP, 2016, pp. 5115–5119.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 06 , June - 2026

[12] D. Snyder et al., "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in Proc. IEEE ICASSP, 2018, pp. 5329–5333.

[13] A. van den Oord et al., "WaveNet: A Generative Model for Raw Audio," arXiv:1609.03499, Sep. 2016.

[14] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-Based Generative Network for Speech Synthesis," in Proc. IEEE ICASSP, 2019, pp. 3617–3621.

[15] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in NeurIPS, vol. 33, 2020.

[16] X. Wang et al., "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," Computer Speech and Language, vol. 64, 2020.

[17] H. Tak et al., "End-to-End Anti-Spoofing with RawNet2," in Proc. IEEE ICASSP, 2021, pp. 6369–6373.

[18] J. Jung et al., "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks," in Proc. IEEE ICASSP, 2022, pp. 6367–6371.

[19] J. Yi et al., "Half-Truth: A Partially Fake Audio Detection Dataset," in Proc. Interspeech, 2022, pp. 4862–4866.

[20] K. Ito and L. Johnson, "The LJ Speech Dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[21] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," University of Edinburgh, 2017.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in Proc. Interspeech, 2017, pp. 2616–2620.

[23] J. K. Chorowski et al., "Attention-Based Models for Speech Recognition," in NeurIPS, vol. 28, 2015.

[24] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in Proc. Interspeech, 2020, pp. 3830–3834.

[25] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," in Proc. ICML, 2023, pp. 28492–28518.

[26] Y. Skerry-Ryan et al., "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in Proc. ICML, 2018.

[27] K. Wang et al., "MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation," in Proc. ECCV, 2020.

[28] N. Kalchbrenner et al., "Efficient Neural Audio Synthesis," in Proc. ICML, 2018, pp. 2410–2419.

[29] T. Kaneko et al., "iSTFTNet: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform," in Proc. IEEE ICASSP, 2022, pp. 6211–6215.

[30] P. Sanroman et al., "Proactive Detection of Voice Cloning with Localized Watermarking," arXiv:2401.17264, Jan. 2024.

[31] S. Baby et al., "Resources for Indian Languages," in Proc. CBBLR Workshop (Interspeech), 2016.

[32] M. Le et al., "Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale," in NeurIPS, vol. 36, 2023.