RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt26416

Embedding-Guided Neural Voice Conversion for Indian Regional-Language Speech Transformation

A. Bala Raju [a, b, *], S.P. Singh [b], Dhiraj Sunehra [c]

[a] Department of Electronics and Communication Engineering, JNTUH, Kukatpally, Hyderabad, Telangana, India.

[b] Department of Electronics and Communication Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India.

[c] Department of Electronics and Communication Engineering, JNTUH UCER, Rajanna Sircilla, Telangana, India.

* Corresponding Author Email: balaraju_ap@yahoo.co.in

DOI: https://doi.org/10.54392/irjmt26416

Received: 21-01-2026; Revised: 07-07-2026; Accepted: 13-07-2026; Published: 24-07-2026

Abstract: Voice conversion (VC) is an emerging technology in speech processing that aims to modify an utterance so that it sounds as though it were spoken by another speaker while preserving its linguistic content. High-quality voice conversion has broad applications, including speech synthesis, assistive communication, and entertainment applications such as multilingual dubbing. However, current embedding-guided voice conversion (EGVC) frameworks often struggle with generalization and naturalness under regional data-scarcity conditions. This study explores these limitations by evaluating an EGNVC framework adapted for low-resource regional-language pairs. The proposed framework incorporates the Harvest pitch-extraction algorithm alongside pretrained speaker representations to guide cross-gender pitch transitions while attempting to preserve speaker-identity profiles. Experimental results show that, although the framework successfully shifts macro-level pitch contours across genders, spectral alterations lead to substantial acoustic distortion and reduced intelligibility. Specifically, Kannada speech conversion achieved a localized objective intelligibility score of STOI = 0.12, whereas Malayalam transformations exhibited substantial spectral variation, with an MCD of 169.75, highlighting significant language-specific barriers to regional voice conversion. Kannada achieved higher intelligibility (STOI = 0.12) than Malayalam, whereas Malayalam required greater spectral modification (MCD = 169.75), indicating language-specific challenges in voice conversion. These baseline metrics delineate the empirical limitations of current embedding-guided architectures for Dravidian languages and indicate that substantial advances in spectral mapping are required before such systems can be integrated into real-time assistive or localized voice-synthesis applications.

Keywords: Voice Conversion, Speaker Embeddings, Neural Voice Transformation, Indian Regional Languages, Generative Adversarial Networks (GANs), Pitch Extraction.

1. Introduction

Voice conversion (VC) is an emerging area of speech processing that enables the voice of one speaker to be transformed into that of another while preserving the linguistic content. This capability has a wide range of applications, including personalized speech synthesis, voice rehabilitation, dubbing, and entertainment [1]. Among the various VC approaches, embedding-guided voice conversion has attracted considerable attention because it can exploit pretrained speaker embeddings for effective and flexible speaker adaptation. Unlike conventional voice-conversion methods that rely on paired training data, embedding-guided methods can perform cross-speaker transformations without explicit source-target alignment. Nevertheless, systems in this category still suffer from significant limitations that restrict their practical utility [2].

One challenge associated with current embedding-guided VC approaches is their poor generalization across speakers and conditions. Most models perform well on their training and test sets but fail to maintain comparable performance for unfamiliar voices, acoustic conditions, or speaking styles [3]. This limited flexibility makes them less suitable for general applications, in which a voice-conversion system should function reliably for many users with different voice types [4]. The dependence on substantial speaker variability in the training data further exacerbates the problem because large-scale datasets are costly and time-consuming to collect.

A further major limitation is the computational inefficiency of existing methods. Many recent VC models are based on complex neural architectures that require substantial computing resources and are unsuitable for

real-time applications [5]. They often require high-end hardware and therefore cannot be readily deployed for live communication, assistive technologies, or mobile applications [6]. In addition, training such models typically involves lengthy and resource-intensive optimization, which further limits scalability and widespread adoption.

In addition to computational limitations, the quality and naturalness of converted voices produced by previous systems remain inadequate. Although recent deep-learning techniques have made synthesized speech increasingly intelligible, artifacts, unnatural prosody, and speaker inconsistency remain perceptible [7]. These artifacts degrade the perceptual quality of converted speech and reduce user satisfaction, particularly in applications requiring high fidelity and expressiveness, such as virtual assistants, audiobook narration, and content localization. Therefore, a framework that improves both the efficiency and perceptual quality of embedding-guided voice conversion is needed [8].

To address these limitations, this paper proposes the Embedding-Guided Neural Voice Conversion (EGNVC) framework, a new method for on-the-fly voice conversion between speakers without requiring large-scale paired training data for every source-target speaker pair. The novelty of EGNVC lies in its adaptive use of pretrained speaker embeddings, which improves generalization across speakers and conditions. This approach reduces dependence on large datasets while maintaining conversion accuracy, thereby increasing its practical applicability. EGNVC is designed to provide high-quality voice conversion with low latency and computational cost through the use of efficient neural-network architectures.

This study primarily contributes an adaptation of an embedding-guided GAN-based voice-conversion pipeline, based on RVC v2, for Indian regional languages. The framework incorporates 768-dimensional speech-feature extraction, RMVPE-based F0 conditioning, FAISS-based speaker-feature retrieval, and GAN-based waveform synthesis for gender-dependent voice conversion in Telugu, Tamil, Malayalam, and Kannada. The remainder of the paper is organized as follows. Section 2 reviews state-of-the-art voice-conversion models. Section 3 describes the proposed model, and Section 4 presents and discusses the experimental results.

2. Related Work

Table 1 presents recent developments in voice-conversion systems and highlights their relevance to the proposed EGNVC framework. It provides a comprehensive overview of current voice-conversion techniques, speaker- and pitch-conditioning strategies, target datasets, and major limitations identified in existing studies. The table organizes the principal methods, research gaps, and contributions, enabling a clear comparison between previous approaches and the proposed regional-language gender voice-conversion framework.

3. Methodology

This section introduces the proposed Embedding-Guided Neural Voice Conversion (EGNVC) framework, which incorporates the Harvest pitch-extraction algorithm for high-quality and dynamic voice conversion. The principal motivation for this approach is that it does not require large paired datasets and can be readily adapted to arbitrary speakers. By combining speaker embeddings with accurate pitch extraction, the model is designed to be scalable, natural, and practical for diverse voice-conversion tasks. In the proposed EGNVC model, preprocessing involves resampling the input speech to 32 kHz and segmenting it. The RVC v2 feature extractor is then used to derive 768-dimensional acoustic-content features.

During training, the RMVPE-GPU method is used to extract F0 information. The proposed architecture ultimately performs F0-guided conversion; therefore, both coarse and normalized F0 features are incorporated into model training.

Figure 1 illustrates the content encoder and prior-network architecture of the proposed EGNVC model. The raw source audio, sampled at 16 kHz, is first processed by the ContentVec/HuBERT feature extractor, which uses convolutional and transformer layers to extract linguistic-content features while reducing speaker-dependent information. These features are then passed through a linear projection layer and a FAISS indexing matcher, which retrieves target-speaker-related feature representations through k-nearest-neighbour matching. Finally, the prior-encoder projection estimates the prior mean and variance, and the normalizing-flow module refines the latent representation to generate the final prior vector for voice conversion.

Figure 2 presents the posterior-encoder architecture of the proposed EGNVC model, which converts the target speaker's linear spectrogram into a latent acoustic representation. The target linear spectrogram, containing 513 frequency bins, is first passed through a one-dimensional convolutional layer that projects the feature dimension from 513 to 192 channels. These features are then processed by a 16-block non-causal WaveNet stack with dilated convolutions, gated activations, and residual-skip connections to capture detailed temporal and spectral patterns. Finally, an output-projection layer generates the mean and log-variance vectors, which are passed through the reparameterization operation to produce the latent vector z used for voice conversion.

Table 1. Comparison of Recent Voice Conversion Systems and Proposed EGNVC

Study	Objective	Methodology	Key Findings
Shashwat Aggarwal *et al.*, 2025 [9]	Develop an end-to-end audio style transformation framework.	Proposed a differentiable architecture for direct audio-to-audio transformation. Supports voice conversion and musical style transfer while preserving content.	Achieved effective style transfer with improved content preservation. Generated high-quality converted speech and music styles.
Chae-Woon Bang *et al.*, 2023 [10]	Improve zero-shot multi-speaker text-to-speech synthesis.	Used information perturbation and a speaker encoder to enhance speaker representation learning.	Generated natural speech for unseen speakers and improved speaker similarity.
Wushour Slam *et al.,* 2023 [11]	Review low-resource speech recognition techniques.	Surveyed recent speech recognition methods, datasets, and deep learning approaches for low-resource languages.	Identified key challenges and highlighted promising directions for improving recognition accuracy.
Xuexin Xu *et al.*, 2023 [12]	Develop an any-to-any voice conversion system.	Introduced multi-layer speaker adaptation with content supervision to separate speaker and linguistic information.	Improved speaker similarity and speech quality across multiple source-target speaker pairs.
Ananya Angra *et al.,* 2024 [13]	Improve spoken dialect identification performance.	Combined aggregated Wav2Vec 2.0 features with a dual-stream TDNN architecture.	Achieved efficient dialect classification and outperformed baseline approaches.
Kadria Ezzine *et al.,* 2023 [14]	Develop a non-parallel any-to-one voice conversion system.	Used an autoregressive conversion model integrated with the LPCNet vocoder.	Generated natural converted speech without requiring parallel training data.
Wen-Shin Hsu *et al.,* 2024 [15]	Enhance dysarthric voice conversion.	Integrated Fuzzy Expectation Maximization with diffusion models for phoneme prediction and speech conversion.	Improved phoneme prediction and speech intelligibility for dysarthric speakers.
Dongsuk Yook *et al.,* 2024 [16]	Improve non-parallel voice conversion quality.	Proposed CycleDiffusion using cycle-consistent diffusion models and cycle-consistency loss.	Produced more natural converted speech and improved conversion performance.
Kun Zhou *et al.,* 2022 [17]	Control emotion intensity in emotional voice conversion.	Developed a framework to model and manipulate emotion intensity during voice conversion.	Enabled flexible emotion control while maintaining speaker identity and speech quality.
Deepak Suresh Asudani *et al.,* 2023 [18]	Review word embedding models for text analytics.	Analyzed traditional and contextual embeddings and their use in deep learning applications.	Found contextual embeddings significantly improve NLP task performance.
Marco Matassoni *et al.,* 2024 [19]	Improve speaker anonymization for voice conversion.	Disentangled speaker information from pre-trained speech embeddings before conversion.	Enhanced privacy protection while maintaining speech intelligibility and naturalness.
Madiha Amarjouf *et al.,* 2024 [20]	Evaluate self-supervised denoising methods for esophageal speech enhancement.	Compared multiple self-supervised denoising techniques on esophageal speech datasets.	Improved speech quality and intelligibility over conventional enhancement methods.
Zhang *et al.,* 2024 [21]	Improve cross-lingual voice conversion while preserving speaker identity.	Proposed RefXVC using global and local speaker embeddings, timbre encoder, pronunciation matching network, and multiple reference speeches.	Improved speech quality and speaker similarity, outperforming existing cross-lingual voice conversion methods.
Guo *et al.,* 2025 [22]	Improve any-to-any voice conversion by	Proposed IAFF-VC with an encoder-decoder architecture and	Achieved state-of-the-art performance with improved

	balancing content preservation and speaker similarity.	EMAFF-based attentional feature fusion for multi-scale speech feature integration.	speaker similarity, speech naturalness, and content accuracy in one-shot voice conversion.
Yamamoto *et al.,* 2023 [23]	Improve singing voice conversion with limited target speaker data.	Proposed a diffusion-based any-to-any voice conversion model trained on large-scale speech and singing datasets and fine-tuned for target speakers.	Achieved competitive naturalness and speaker similarity, demonstrating strong generalization in cross-domain singing voice conversion.
Sisman *et al.,* 2020 [24]	Review the evolution of voice conversion techniques and their challenges.	Surveyed voice conversion approaches from statistical methods to deep learning-based models, including feature extraction, speaker representation, and conversion frameworks.	Provided a comprehensive overview of voice conversion technologies, highlighting recent advances, challenges, and future research directions.
Chen *et al.,* 2024 [25]	Improve singing voice conversion efficiency while maintaining sound quality and timbre similarity.	Proposed LCM-SVC, a latent diffusion model accelerated using latent consistency distillation (LCD) for one-step or few-step inference.	Significantly reduced inference time while preserving high sound quality and timbre similarity, outperforming existing SVC methods in efficiency.
Lu *et al.,* 2022 [26]	Develop a one-shot emotional voice conversion system for seen and unseen emotions.	Proposed a feature separation framework using Global Emotion Embeddings (GEEs), AdaIN, Activation Guidance (AG), and Mutual Information (MI) loss.	Successfully converted seen and unseen emotions without emotion labels, achieving effective emotion transfer and feature disentanglement.
Kameoka *et al.*, 2025 [27]	Improve non-parallel voice conversion quality and inference speed.	Proposed LatentVoiceGrad, integrating latent diffusion and flow-matching models into the VoiceGrad framework.	Enhanced speech quality and significantly accelerated voice conversion compared to the original VoiceGrad method.
Guo *et al.*, 2023 [28]	Improve cross-lingual and expressive voice conversion performance.	Proposed a joint training speaker encoder with speaker consistency loss and Whisper-based content features.	Enhanced speaker similarity and effectively preserved speaker characteristics and emotional expressions in converted speech.
Qian *et al.,* 2019 [29]	Zero-shot voice conversion using speaker-style transfer without parallel speech data	Developed AutoVC using an autoencoder bottleneck with speaker embedding conditioning	Achieved voice conversion using only reconstruction loss, but mainly focused on general voice conversion and not Indian regional languages
Li *et al.,* 2021 [30]	Diverse non-parallel many-to-many voice conversion with natural-sounding speech	Used StarGANv2-VC with GAN architecture, style encoder, mapping network, and adversarial learning	Produced diverse and natural voice conversion, but required careful adversarial training and did not target low-resource Indian languages
Das *et al.,* 2023 [31]	Emotion-preserving voice conversion using an improved StarGAN-based model	Applied StarGAN-VC++ with deep speaker and emotion embeddings	Improved emotional consistency in converted speech, but focused mainly on emotion preservation rather than regional-language gender conversion
RVC-Project, 2023 [32]	Efficient retrieval-based voice conversion with limited training data	Used self-supervised speech features, FAISS retrieval, and F0/pitch conditioning	Provided strong practical voice conversion performance, but formal evaluation for Indian regional languages is limited

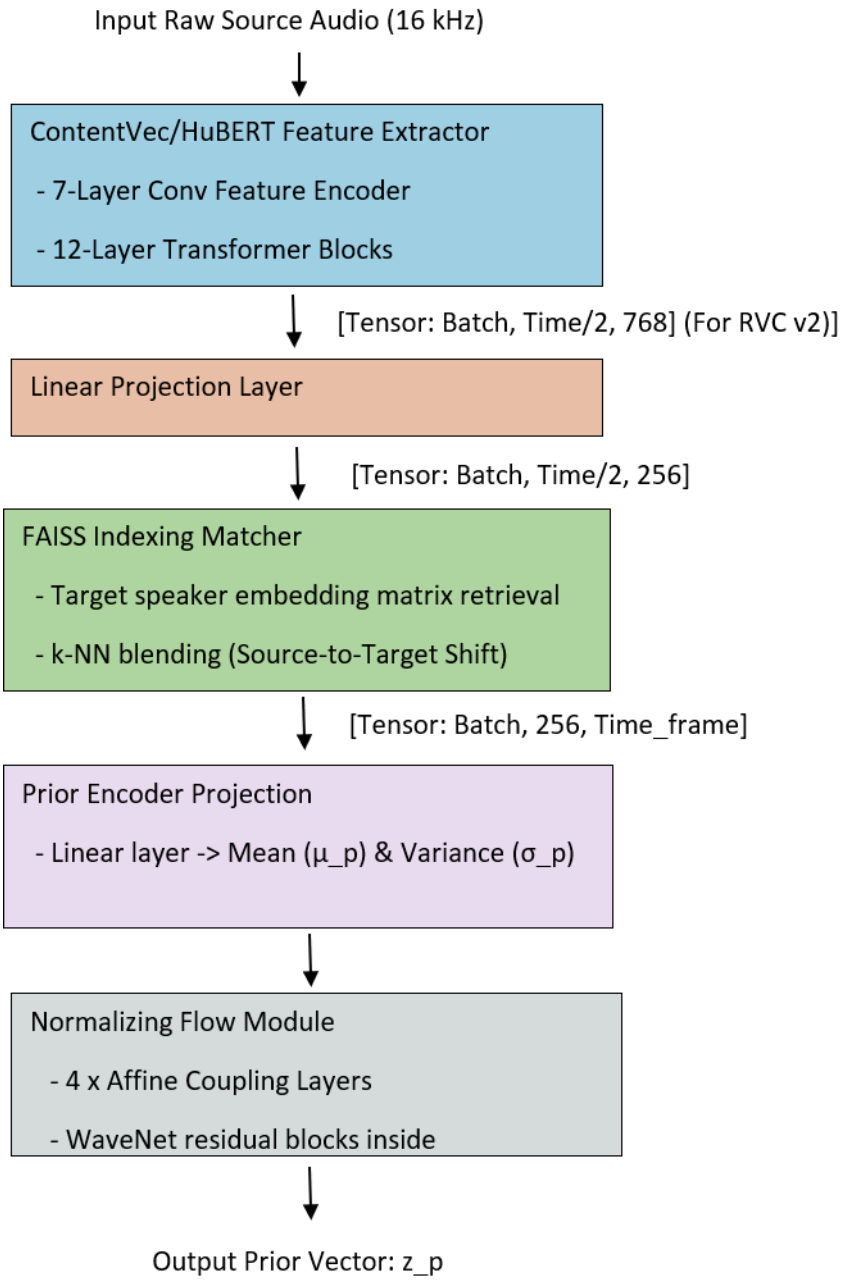

Figure 1. Content Encoder and Prior Network Architecture of the Proposed EGNVC Model.

Figure 3 shows the decoder architecture with the NSF-HiFiGAN generator, which synthesizes the final converted speech waveform from the latent vector and pitch information. The F0 vector is processed by the RMVPE pitch estimator and converted into a sine-wave excitation using the NSF harmonic oscillator, while the latent vector z is passed through a pre-convolution layer. These two representations are combined and passed through upsampling convolutional blocks, where temporal resolution is increased using transposed one-dimensional convolutions. Multi-receptive-field residual blocks with different kernel sizes capture both short- and long-term speech patterns, and their outputs are summed before further upsampling. Finally, a one-dimensional convolutional layer with a tanh activation generates the converted output waveform with improved naturalness and speaker similarity.

Figure 4 illustrates the dual-discriminator configuration used in the proposed EGNVC model to improve the realism and naturalness of the generated speech. The raw input waveform is evaluated through two discriminator branches: the Multi-Period Discriminator (MPD), which analyzes periodic speech patterns using periods of 2, 3, 5, 7, and 11, and the Multi-Scale Discriminator (MSD), which examines the waveform at its original scale and after 2× and 4× downsampling. The MPD reshapes the one-dimensional waveform into two-dimensional representations and passes them through multiple convolutional layers to capture harmonic and temporal structures. Finally, the

classifier-projection head validates the feature distributions and produces a real/fake scalar output, thereby helping the generator synthesize more natural and higher-quality speech.

The proposed model is trained using pretrained RVC v2 generator and discriminator checkpoints. The generator transforms linguistic-content features and F0 information into an acoustic representation of the target speaker, while the discriminator provides adversarial feedback to enhance the naturalness of the converted speech. A FAISS-based retrieval index is constructed from the extracted 768-dimensional features using an IVF-Flat index with nprobe set to 1. In the final stage, the RVC v2 GAN-based neural-vocoder pipeline generates the output waveform. During inference, the trained model checkpoint and FAISS index are used with an index-retrieval ratio of 0.5 and a consonant-protection value of 0.5. The model flowchart is shown in Figure 5.

In the EGNVC framework shown in Figure 5, Harvest is used to extract pitch from the source audio, after which the pitch contour is adjusted to align with the target speaker's pitch characteristics.

This process helps the output audio sound natural and closely resemble the target speaker's vocal style. Let the source speech waveform be denoted by

$$x_s(t), t = 1,2, \dots, T \quad (1)$$

Where T represents the number of audio samples. The objective is to generate a converted waveform

$$\hat{x}_t(t) \quad (2)$$

Such that the linguistic content of $x_s(t)$is preserved while the speaker identity matches a target speaker characterized by a reference utterance$x_t(t)$.

The conversion process can therefore be expressed as

$$\hat{x}_t = \mathcal{F}(x_s, x_t) \quad (3)$$

Where $\mathcal{F}(\cdot)$ denotes the proposed EGNVC transformation model.

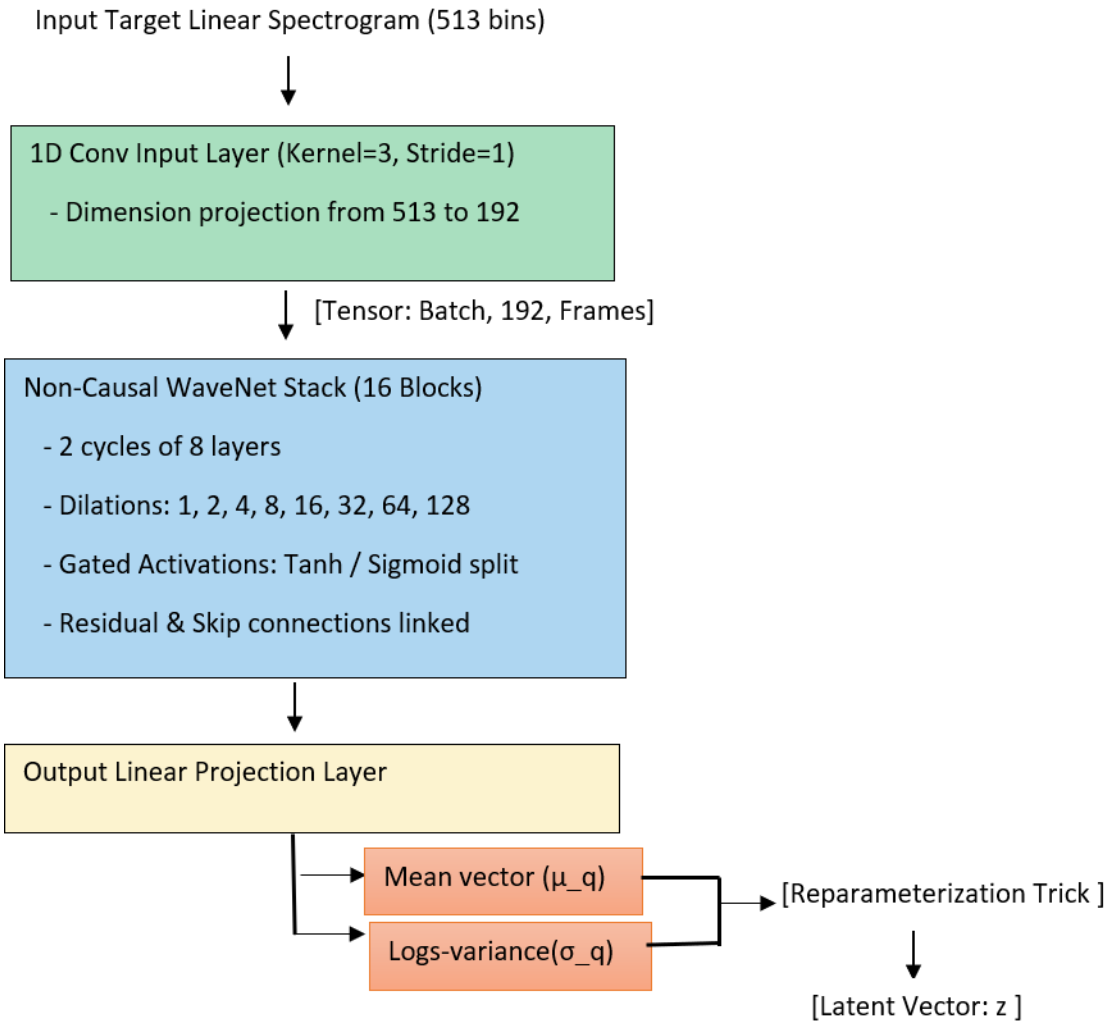

Figure 2. Posterior Encoder Architecture.

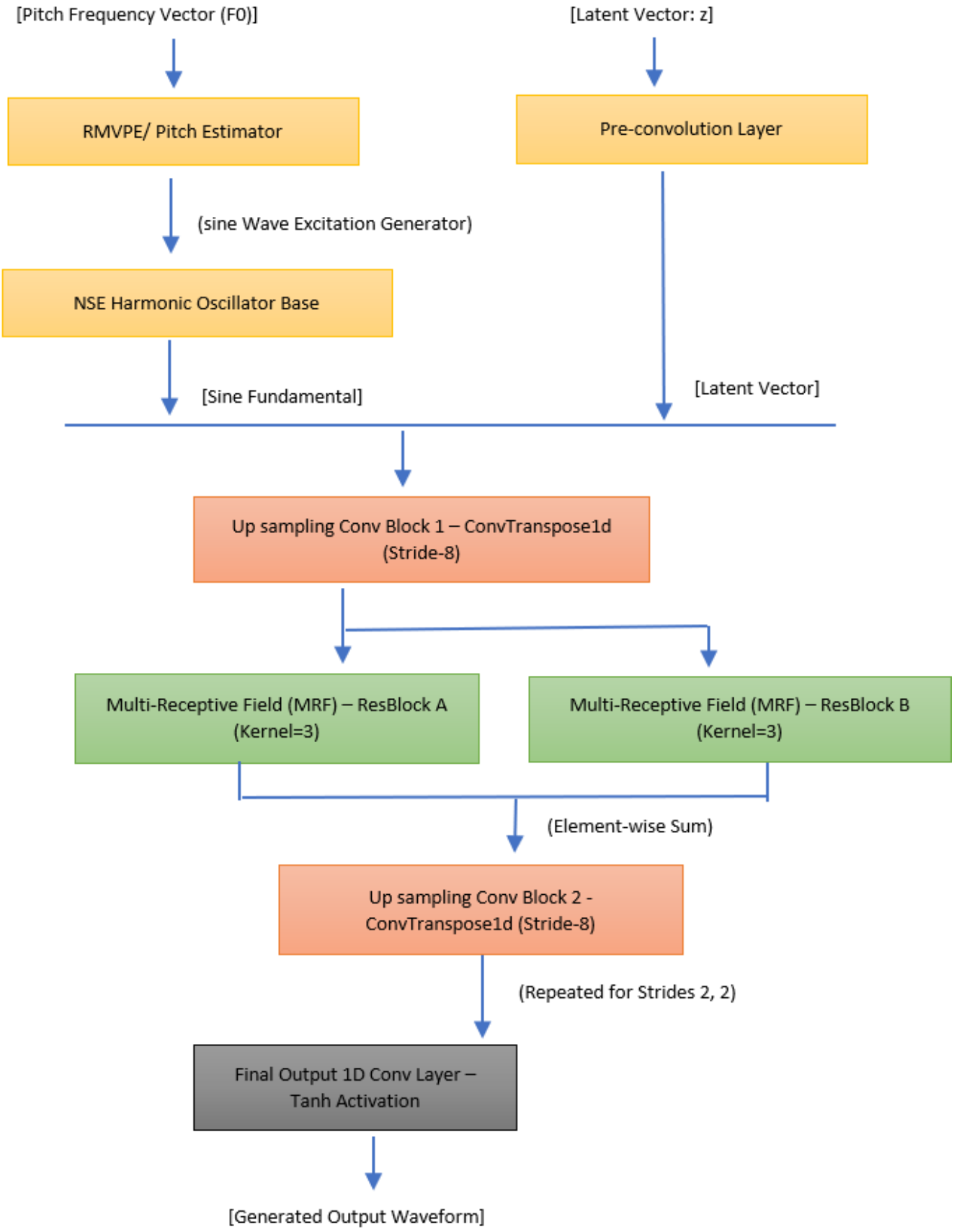

Figure 3. Decoder Architecture with NSF-HiFiGAN Generator.

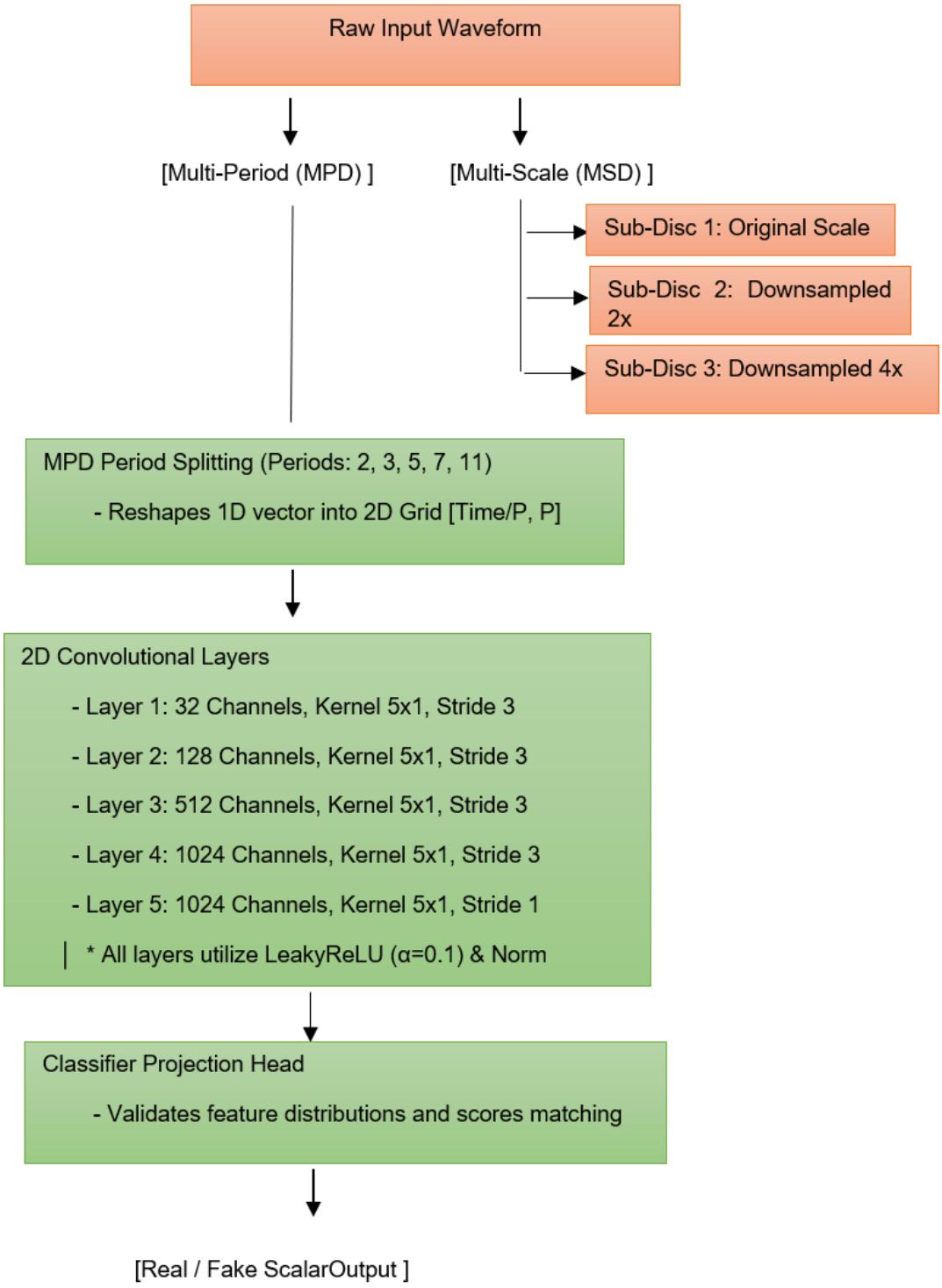

Figure 4. Dual Discriminator Configuration.

3.1 Acoustic Preprocessing

The source waveform is first resampled to 32 kHz and segmented into short-time overlapping frames. For each frame, a Short-Time Fourier Transform (STFT) is computed as

$$X(m,k) = \sum_{n=0}^{N-1} x_s\,(n + mH)w(n)e^{-\frac{j2\pi kn}{N}} \quad (4)$$

Where:

- N= frame length
- H= hop size
- $w(n)$= analysis window
- m= frame index
- k= frequency bin index

The magnitude spectrum is projected onto the mel scale to obtain the log-mel spectrogram:

$$M_s = \log\,(\text{Mel}(|\,X\,|)) \quad (5)$$

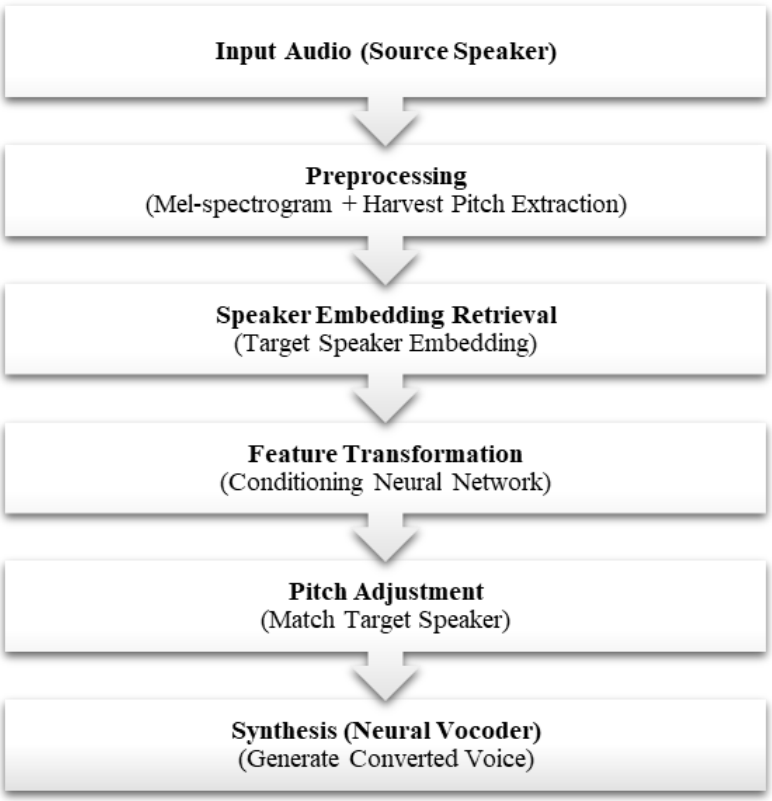

Figure 5. Flowchart of the Proposed EGNVC model.

Where

$$M_s \in R^{F \times L} \quad (6)$$

with Fmel-frequency bins and Ltime frames.

The log-mel spectrogram provides a compact, perceptually meaningful representation for subsequent conversion.

3.2 Pitch Extraction and F0 Normalization

To preserve natural prosody while adapting speaker characteristics, frame-wise fundamental frequency (F0) is estimated using the Harvest pitch estimator:

$$f_s(l), l = 1,2, \dots, L \quad (7)$$

Where $f_s(l)$is the source pitch in Hertz.

Only voiced frames are used for statistical normalization. Let V denote the set of voiced frames where$f_s(l) > 0$. The source pitch statistics are:

$$\mu_s = \frac{1}{|V|} \sum_{l \in V} \log f_s(l) \quad (8)$$

$$\sigma_s = \sqrt{\frac{1}{|V|} \sum_{l \in V} (\log f_s(l) - \mu_s)^2} \quad (9)$$

Similarly, the target speaker pitch statistics are denoted by $\mu_t$andσ_t.

The source pitch is mapped into the target speaker pitch space using mean-variance normalization:

$$\log \hat{f}_t(l) = \mu_t + \frac{\sigma_t}{\sigma_s} (\log f_s(l) - \mu_s) \quad (10)$$

Thus,

$$\hat{f}_t(l) = \exp (\log \hat{f}_t(l)) \quad (11)$$

This transformation preserves pitch-contour dynamics while adapting the average pitch range and variance to those of the target speaker.

3.3 Speaker Embedding Conditioning

Speaker identity is represented using a pretrained speaker encoder that extracts a fixed-dimensional embedding vector:

$$e_t \in \mathrm{R}^d \quad (12)$$

Where $d = 256$in the final implementation.

This embedding captures timbre, vocal-tract characteristics, and speaker style. Unlike one-to-one conversion systems, the embedding enables flexible conditioning without explicit paired-speaker mapping.

For each acoustic frame, the encoder network produces latent content features:

$$h_l = Enc(M_s(:, l)) \quad (13)$$

The conditioned feature vector is formed as

$$z_l = [h_l; e_t; \hat{f}_t(l)] \quad (14)$$

Where [;]denotes vector concatenation.

Thus, each frame is jointly conditioned on:

- source linguistic content
- target speaker identity
- converted pitch contour

3.4 Acoustic Feature Conversion Network

The generator network transforms conditioned features into converted acoustic features:

$$\widehat{M}_t = G(z_1, z_2, \dots, z_L) \quad (15)$$

Where $\widehat{M}_t$ is the predicted target mel-spectrogram.

In the practical implementation, the conversion backbone follows an RVC-v2-style generator with adversarial refinement. The generator learns spectral adaptation, temporal consistency, and speaker-dependent timbre conversion.

3.5 Neural Vocoder Synthesis

The converted mel representation is transformed into time-domain waveform samples using a GAN-based neural vocoder:

$$\hat{x}_t = Voc(\widehat{M}_t) \quad (16)$$

where $Voc(\cdot)$denotes the waveform synthesizer.

This stage reconstructs natural speech with improved phase continuity, intelligibility, and perceptual realism.

3.6 Training Objective Function

The network parameters are optimized using a multi-objective loss:

$$L_{total} = \lambda_1 L_{mel} + \lambda_2 L_{adv} + \lambda_3 L_{fm} + \lambda_4 L_{spk} + \lambda_5 L_{f0} \quad (17)$$

where:

(a) Mel Reconstruction Loss

$$L_{mel} = || M_t - \widehat{M}_t ||_1 \quad (18)$$

This term enforces spectral similarity between the generated and reference speech.

(b) Adversarial Loss

$$L_{adv} = \mathbb{E}[\log D(x_t)] + \mathbb{E}[\log (1 - D(\hat{x}_t))] \quad (19)$$

where $D(\cdot)$is the discriminator.

(c) Feature Matching Loss

$$L_{fm} = \sum_r || D_r(x_t) - D_r(\hat{x}_t) ||_1 \quad (20)$$

where $D_r(\cdot)$denotes intermediate discriminator layer features.

(d) Speaker Consistency Loss

$$L_{spk} = 1 - \cos (E(x_t), E(\hat{x}_t)) \quad (21)$$

where $E(\cdot)$is the speaker encoder.

(e) Pitch Consistency Loss

$$L_{f0} = || f_t - \hat{f}_t ||_1 \quad (22)$$

Empirically selected weights are:

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = (1.0, 0.5, 2.0, 1.0, 1.0) \quad (23)$$

Algorithm: EGNVC Voice Conversion Pipeline

Input:

Source speech $x_s(t)$

Target speaker reference $x_t(t)$

Output:

Converted speech $\hat{x}_t(t)$

1. Extract mel-spectrogram M_s from $x_s(t)$
2. Estimate source *F0* f_s using Harvest
3. Extract target speaker embedding e_t
4. Compute source and target pitch statistics
5. Convert source pitch using mean-variance mapping
6. Encode source mel features
7. Concatenate encoded features + e_t + converted *F0*
8. Generate converted mel spectrogram $\widehat{M}_t$
9. Use neural vocoder to synthesize waveform $\hat{x}_t(t)$
10. Return converted speech

4. Result Analysis

The experimental results obtained with the implemented EGNVC framework are presented in this section. The proposed EGNVC model was evaluated in terms of speaker similarity, speech quality, and conversion accuracy. When the Harvest pitch-estimation algorithm was applied, the model successfully shifted the source speaker's timbre while preserving the target speaker's pitch characteristics. The use of pretrained speaker embeddings enabled adaptation to unseen speakers through zero-shot learning, even with limited training data. The experiments used input speech signals from several speakers to assess flexibility and generalization. The dataset contained male and female speech signals that were preprocessed and analyzed before being used to train and test the system. The configuration of the proposed EGNVC system is reported in Table 2.

Figure 6 shows the male speech signal used for training. The signal contains 15,537,152 samples and was recorded at a sampling frequency of 44,100 Hz. The male speech provides a reference for speaker characteristics and the spectral envelope that the model

learns to reproduce or modify during target-speaker adaptation, including cross-speaker synthesis. Figure 7 shows the input female speech signal supplied to the model. This speech sample contains 1,572,853 samples and was also recorded at a sampling rate of 44,100 Hz. The female speech signal is treated as the source signal to be transformed during voice conversion. The model uses the spectral and temporal features of this signal to synthesize an output that resembles the target male speaker while preserving the linguistic content.

Figure 8 shows the output speech converted from a female voice to a male voice. The output contains 1,869,440 samples and is sampled at 32 kHz.

Table 2. Configuration of the Proposed EGNVC System

Parameter	Final Setting Used
Base framework	RVC v2 voice conversion pipeline
Input audio sampling rate	32 kHz
Training audio type	Clean vocal speech
No. of speakers per language	2
No. of languages	4 (Telugu, Tamil, Malayalam and Kannada)
No. of utterances per speaker	6
Recommended audio duration	3–7 minutes per speaker
Acoustic feature dimension	768-dimensional RVC v2 features
F0 extraction during training	RMVPE-GPU
F0-guided training	Yes
Generator pretrained model	f0Ov2Super32kG.pth
Discriminator pretrained model	f0Ov2Super32kD.pth
Batch size	8
Total epochs	35
Checkpoint saving frequency	Every 50 epochs
GPU device	CUDA GPU 0
Feature index	FAISS IVF-Flat
Index probe value	nprobe = 1
Inference F0 method	RMVPE
Index retrieval ratio	0.5
Consonant protection	0.5
Final vocoder	RVC v2 GAN-based neural vocoder

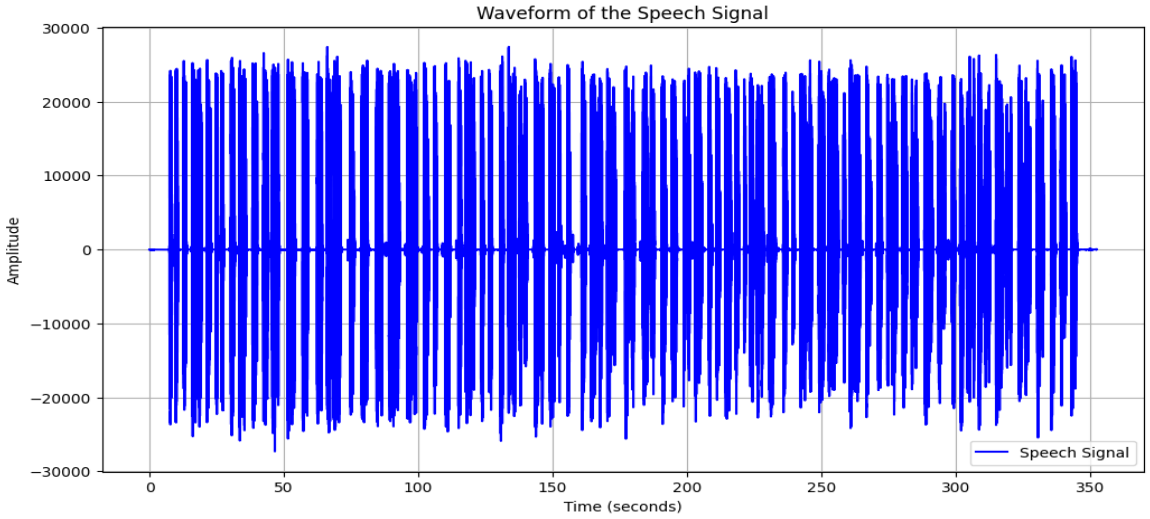

Figure 6. Sample Input Male-Speech Spectrum Used to Train the Model.

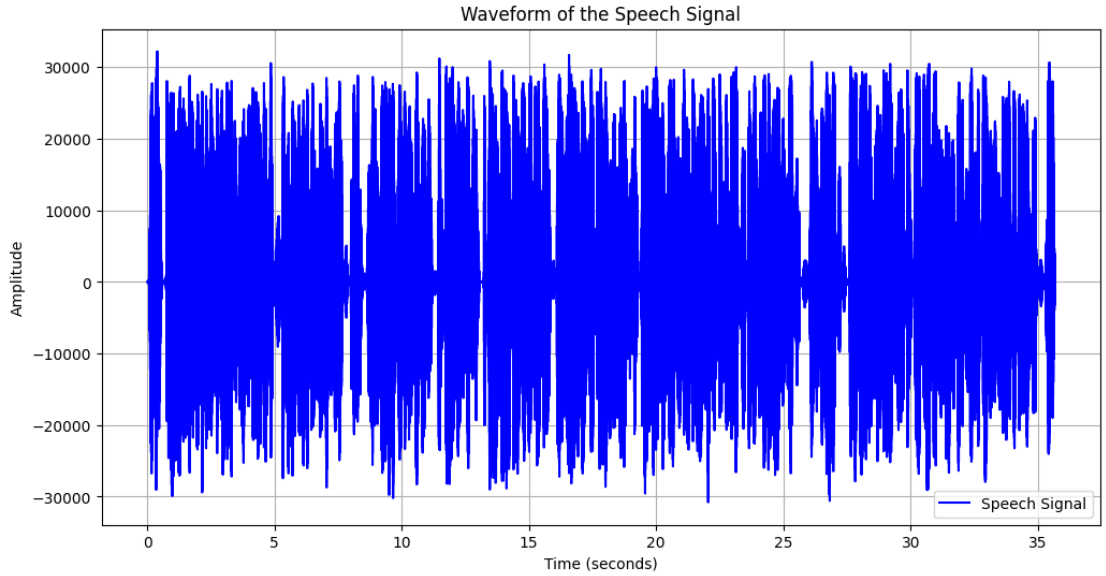

Figure 7. Input Female speech spectrum.

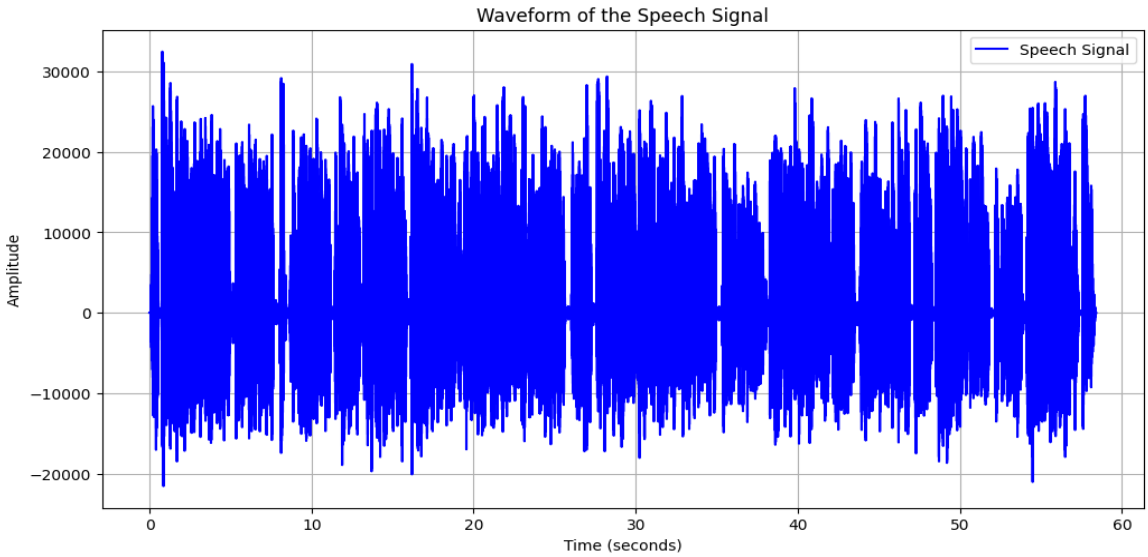

Figure 8. Converted Female-to-Male Speech Spectrum.

The change in sampling rate reflects the model's output configuration for efficient encoding and processing while preserving speech intelligibility and quality. The converted speech retains the linguistic and prosodic elements of the original female voice but alters pitch, timbre, and other speaker-specific features to resemble the male speaker. The model's performance was monitored during training using several loss functions. The discriminator loss (Loss_disc) assesses how effectively the discriminator distinguishes real audio from generated audio, while the generator loss (Loss_gen) evaluates the quality of the generated content. Loss_fm compares intermediate feature representations to improve speech-synthesis quality. The mel-spectrogram loss (Loss_mel) preserves spectral fidelity through mel-spectrogram reconstruction, thereby retaining information about the original structure of the generated speech. In addition, the Kullback-Leibler divergence loss (Loss_kl) regularizes the latent-space distribution by reducing divergence between the predicted and reference distributions.

Discriminator Loss: In GANs, discriminator loss is used to evaluate how effectively the discriminator distinguishes real samples from generated samples.

$$L_D = -E_{x\sim p_{data}}[\log D(x)] - E_{z\sim p_z}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (24)$$

$D(x)G(z)p_{data}p_z$Here, the first term represents the probability assigned by the discriminator to real samples. The generated sample and the real-data and noise distributions are represented by the corresponding variables, respectively.

In EGNVC, adversarial training is governed by the interaction between the generator and discriminator. The discriminator loss is an important metric for evaluating the discriminator's ability to distinguish real speech samples from generated samples. Figure 9 shows the discriminator loss across training epochs. The

number of epochs is plotted on the x-axis, and the discriminator-loss values are plotted on the y-axis. The curve exhibits early oscillations, which are common in adversarial training as the discriminator adapts to the changing outputs of the generator. Initially, the loss is high because the discriminator can readily distinguish real from generated speech. As training progresses, the loss decreases because the generated speech becomes increasingly realistic and therefore more difficult to distinguish from real speech.

A pronounced spike appears near epoch 18, possibly because a large generator-parameter update temporarily destabilized adversarial training. Such oscillatory behavior is expected in GANs because of the dynamic optimization process. After epoch 20, the discriminator loss begins to plateau, indicating convergence and improved speech-synthesis quality. The subsequent stabilization of the discriminator loss suggests that adversarial learning is becoming more balanced and that voice-conversion performance is improving. This trend is supported by the subjective and objective evaluations, which indicate improvements in speaker similarity, intelligibility, and perceptual speech quality with the proposed method.

Generator Loss: The generator loss indicates how effectively the generator produces samples that the discriminator classifies as real.

Figure 10 shows the generator loss across the training epochs. The number of epochs is plotted on the x-axis, and the generator-loss value is plotted on the y-axis. Initially, the loss decreases to a threshold value while fluctuating because of adversarial training. Overall, however, the generator loss increases as the discriminator becomes more effective at distinguishing real from generated speech. A local decrease around epoch 4 indicates a temporary improvement in the generator's ability to fool the discriminator. Nevertheless, the overall increase suggests that the discriminator is learning faster, making it more difficult for the generator to produce realistic samples. This competitive interaction between the generator and discriminator is expected in GAN-based training, in which the generator continually updates its parameters to outperform the discriminator.

An increasing generator loss usually indicates that the discriminator is becoming more effective at distinguishing real from generated samples, which in turn pressures the generator to improve its outputs. The interaction between the generator and discriminator is therefore crucial for improving the quality of converted speech. The results show that the generator initially struggles, but adversarial training progressively enhances voice-synthesis quality, as reflected in the evaluation metrics used in this study. The proposed method is evaluated using both objective and perceptual measures of the converted speech.

Feature Matching Loss: Feature-matching loss encourages the generator to produce samples with statistical properties similar to those of real samples.

$$L_{FM} = E_{x\sim p_{data}}[f(D(x))] - E_{z\sim p_z}\left[f\left(D(G(z))\right)\right] \quad (25)$$

Figure 11 shows the evolution of the feature-matching loss during training. The number of epochs is plotted on the x-axis, and the loss value is plotted on the y-axis. During the early stages of training, the loss begins at a moderate level and exhibits visible fluctuations, as expected in adversarial learning, during which the generator continually adjusts to changing discriminator feedback. The loss increases around epoch 5, indicating a temporary difficulty in aligning generated speech features with real-speech representations. Such oscillation is typical in adversarial settings, particularly when the discriminator changes rapidly and the generator must catch up.

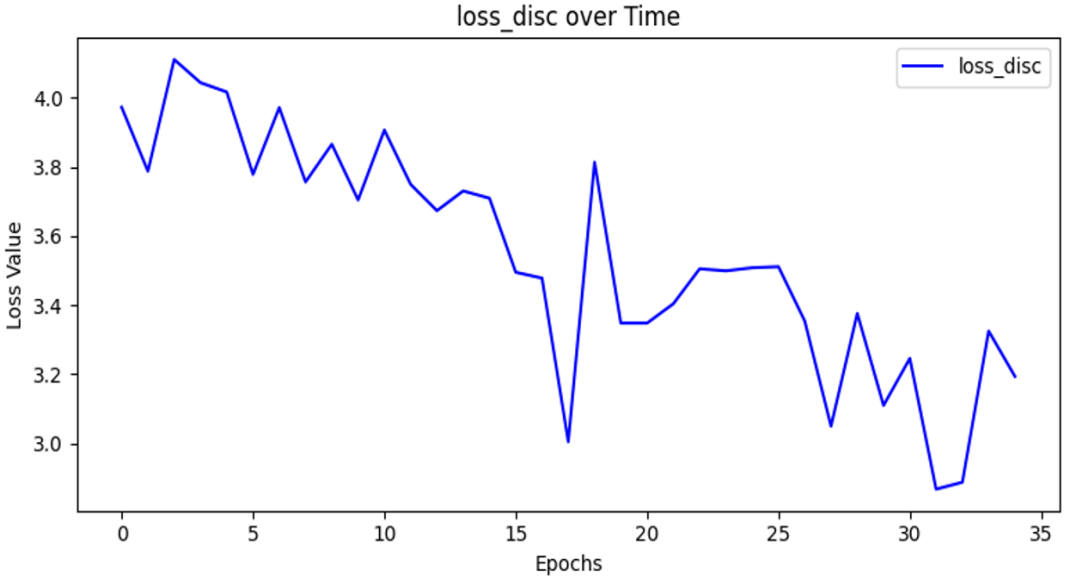

Figure 9. Discriminator Loss versus Training Epoch.

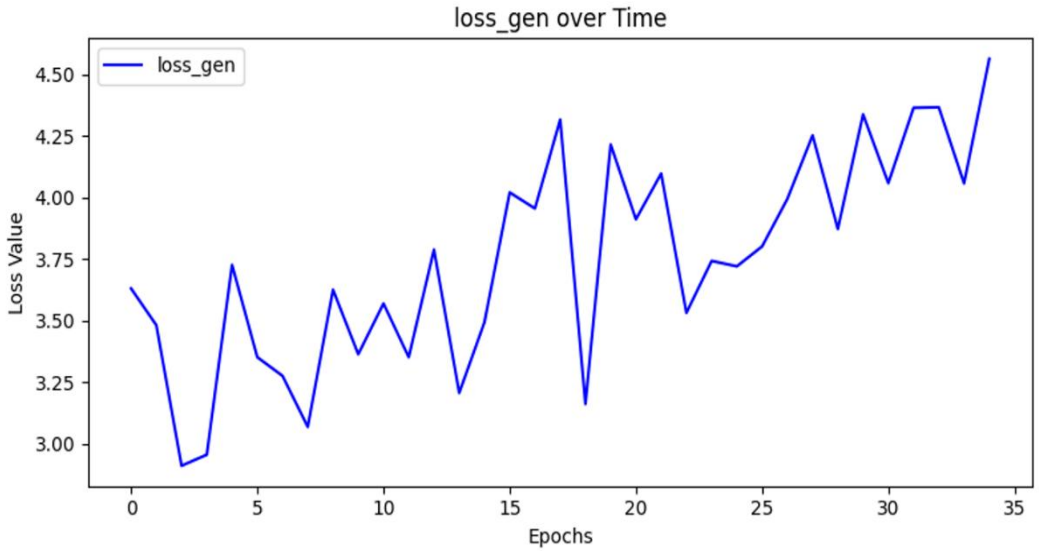

Figure 10. Generator Loss versus Training Epoch.

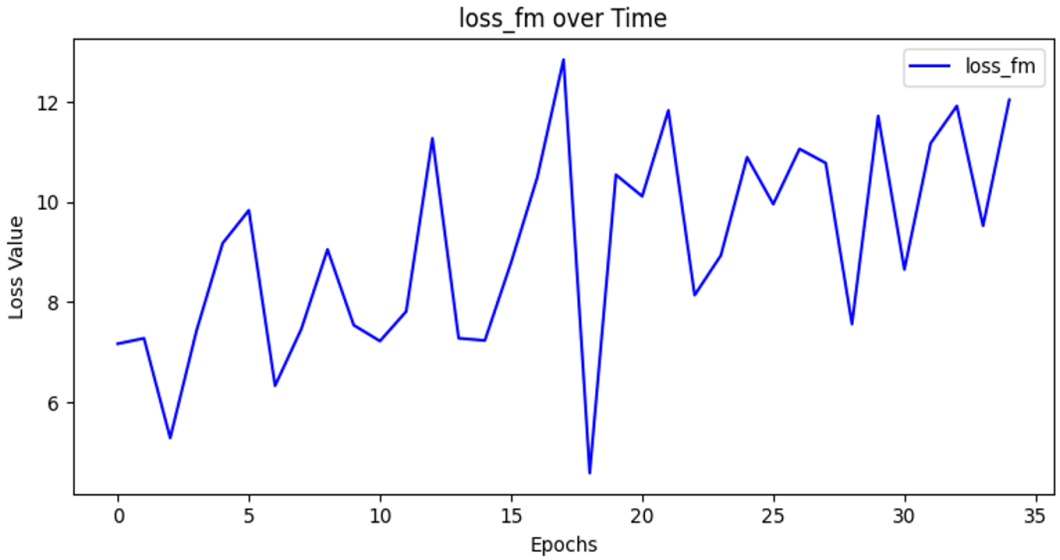

Figure 11. Feature-Matching Loss versus Training Epoch.

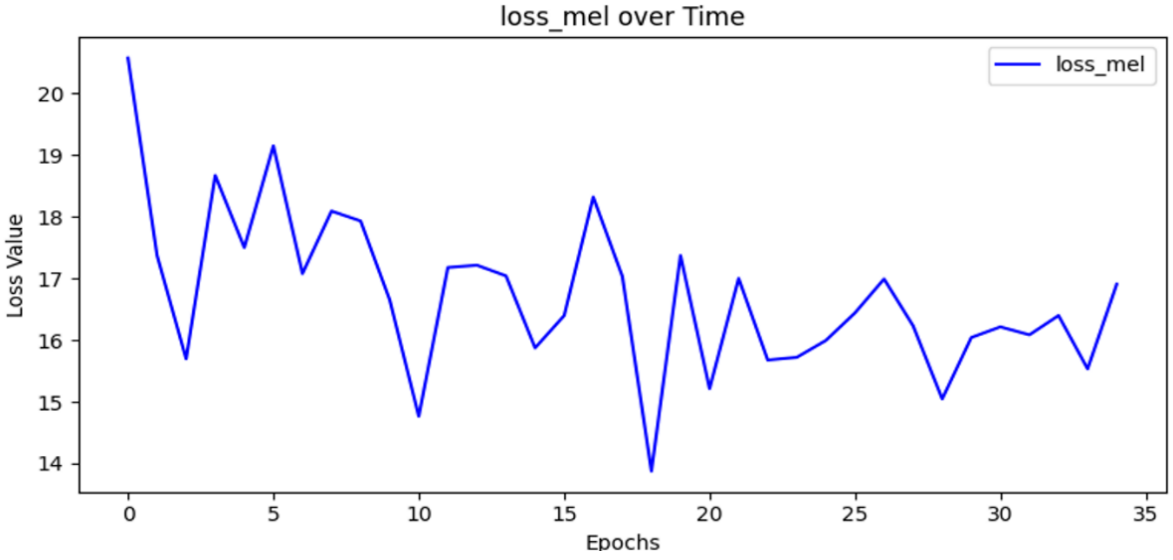

Figure 12. Mel-Spectrogram Loss versus Training Epoch.

The loss continues to fluctuate during training rather than decreasing monotonically. This behavior reflects the adversarial interaction between the generator and discriminator: as the generator's feature representations improve, the discriminator also becomes more effective at detecting residual inconsistencies. A prominent peak occurs near epoch 16, indicating a temporary increase in the separation between real and generated feature distributions and making it more difficult for the generator to produce realistic features.

Despite these fluctuations, the overall training process indicates progressive improvement in the quality of synthesized voice conversion.

Mel-Spectrogram Loss: The mel-spectrogram loss measures the error between the reference and predicted mel-spectrogram representations of speech.

$$L_{Mel} = \|Mel(x) - Mel(\hat{x})\| \quad (26)$$

Figure 12 shows the mel-spectrogram loss across the training epochs. The number of epochs is plotted on the x-axis, and the loss value is plotted on the y-axis. At the beginning of training, the loss is high, at approximately 20, indicating that the model initially has difficulty generating speech with spectral characteristics similar to those of real human speech. During the first few epochs, the loss decreases rapidly, showing that the model quickly learns the basic spectral patterns required for voice conversion.

Despite its early gains, the loss has oscillations during the scale. These changes in the feature vectors are representative of the adversarial learning procedure, with the generator progressively optimizing its own spectral representations over time to better correspond to real speech and the discriminator meanwhile getting stronger at distinguishing between real and generated samples. We find a dramatic decrease in loss around epoch 10, implying that the model briefly converged to a superior spectral alignment. That doesn't look quite right though, but the later fluctuations show that we need another refinement to make it hold more generally across speech samples.

The mel-spectrogram loss plays an important role in improving speech clarity, naturalness, and speaker-identity preservation. A lower mel-spectrogram loss indicates that the generated voice retains finer spectral details required for intelligible and expressive speech. Although the loss oscillations reflect ongoing fine-tuning, the general trend suggests that the model is improving its ability to synthesize realistic speech. The experimental results indicate that incorporating the mel-spectrogram loss into EGNVC training improves the quality of converted speech. The resulting synthetic speech can be evaluated in terms of spectral transitions, pitch control, and speaker similarity, all of which approach the characteristics of natural human speech.

Kullback-Leibler Divergence Loss: Kullback-Leibler (KL) divergence quantifies the difference between two probability distributions.

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (27)$$

where, P and Q represent the true distribution and the approximated one respectively.

Figure 13 shows the KL-divergence loss across the training epochs. The number of epochs is plotted on the x-axis, and the loss value is plotted on the y-axis. At the beginning of training, the KL loss is high, at approximately 9, indicating that the latent representations differ substantially from the desired prior distribution. This is expected during the early stages of model development, when the model is still learning to encode speaker and speech features in a structured representation. As training progresses, the KL loss decreases rapidly during the first few epochs, indicating that the model quickly learns a more coherent latent representation. After this initial decrease, the optimized loss converges to a lower level and fluctuates around it.

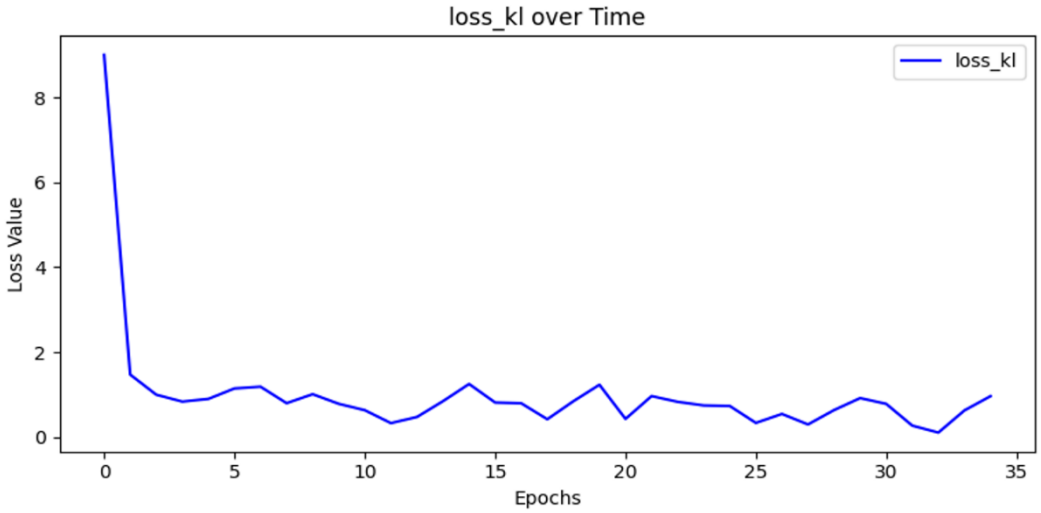

Figure 13. Kullback-Leibler Divergence Loss versus Training Epoch.

The small variations in later epochs reflect continued fine-tuning of the latent space so that the encodings remain interpretable and well separated.

The model's goodness of fit was evaluated using several metrics, including fundamental-frequency (F0) comparison, Short-Time Objective Intelligibility (STOI), and Word Error Rate (WER). F0 comparison measures the extent to which pitch is preserved or appropriately transformed in the converted speech. STOI assesses the intelligibility of the converted speech by comparing its temporal and spectral characteristics with those of reference speech, whereas WER indicates how accurately an automatic speech-recognition system recovers the intended text from the converted speech.

Fundamental Frequency (F0) Comparisons:

Short-Time Objective Intelligibility (STOI): Word Error Rates (WER). F0 Comparison measures the success of pitch retaining in the converted speech. STOI measures how intelligible the converted speech is by comparing the temporal and spectral characteristics of it with those of reference speech, whereas WER is an indication of how well automatic speech recognition (ASR) recognizes the text given to it in response to the converted speech.

Comparison of Fundamental Frequency (F0):

The lowest-frequency component of a periodic waveform is known as the fundamental frequency, or pitch, and is usually expressed in hertz. It is an important feature in speech analysis and voice transformation. F0 is measured using the following relationship:

Comparison of Fundamental Frequency (F0):

The lowest frequency component of a periodic waveform is known as the pitch or fundamental frequency (Fn) usually in Hertz. It is an important feature for speech analysis and voice transformation. F0 is measured by an

$$F_0(t) = \frac{1}{T_0(t)} \quad (28)$$

Where $T\ (t)$ represents the primal period of the speech waveform at time. It is common to computethe F0 values of a reference and a synthesizedspeech signals using:

$$RMSE(F_0) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F_0^{ref}(i) - F_0^{syn}(i)\right)^2} \quad (29)$$

Where $F\ (i)$ presents the fundamental frequency of the reference speech. $F\ (i)$ is the base frequency of speech synthesis and N is total frames.

- Short-Time Objective Intelligibility (STOI): [33] STOI is an objective measure of the intelligibility of processed speech signals. It is well-correlated with human perception of intelligibility. STOI measure The STOI158 is a model-based metric and estimates the intelligibility of two correlated speech signals how to measure its performance?_Count(n(T)) = Count(t2 X4(F)); (13)where[] represents element-wise multiplication.
- Divide the signals into short-time overlapping segments using a Hann window

$$x_k(m) = x(m + kH)w(m), y_k(m) = y(m + kH)w(m) \quad (30)$$

- Where H is the frame shift, w(m) is the window function.
- Calculate the short-time power spectrum of each segment *Note: log normalization of PAM.

$$X_k(f) = \sum_m x_k(m)e^{-j2\pi fm/M} \quad (31)$$

$$Y_k(f) = \sum_m y_k(m)e^{-j2\pi fm/M} \quad (32)$$

Short-Time Objective Intelligibility (STOI): [33] STOI is an objective measure of the intelligibility of processed speech signals and is well correlated with human judgments of intelligibility. It is a model-based metric that estimates the intelligibility of a processed signal relative to a corresponding reference signal.

- Where the M is the FFT size.
- Estimate the correlation of clean and degraded speech is over a time window.

$$STOI(x, y) = \frac{1}{K}\sum_k corr(X_k, Y_k) \quad (33)$$

Where $corr$ (X,Y) is the correlation coefficient of clean and degraded signals in frame k.

Word Error Rate (WER): The Word Error Rate (WER) is the traditional measure for ASR performance. It calculates the disparity between a reference transcript and ASR-generated transcript.WER is calculated as:

$$WER = \frac{S+D+1}{N} \quad (34)$$

Here S is substitutions, D is deletions, I is insertions and N is the total number words in reference transcriptREC.

Word Error Rate (WER): WER is a conventional measure of automatic speech-recognition performance. It quantifies the difference between a reference transcript and the transcript generated by the recognition system. WER is calculated as follows:

Objective speech quality and conversion accuracy measures were used to evaluate the performance of the EGNVC framework. Table 3 shows the results of Telugu-language voice conversion experiments for both female-to-male (F → M) and male-to-female (M → F) transformations, respectively.

Table 3. Performance Metrics of the Proposed EGNVC Model for Telugu Female-to-Male and Male-to-Female Conversion

Metric	Value	
	Female to Male	Male to Female
Fundamental Frequency (F0)	51.61	50.89
Short-Time Objective Intelligibility (STOI)	0.12	0.12
Word Error Rate (WER)	4.24	0.98
Mel Cepstral Distortion (MCD)	276.646	292.819

Here, S, D, and I denote substitutions, deletions, and insertions, respectively, and N is the total number of words in the reference transcript.

The proposed EGNVC system was evaluated across four Indian regional languages—Telugu, Tamil, Kannada, and Malayalam—to assess its performance under diverse phonetic and linguistic conditions. These languages differ substantially in intonation, pronunciation, and prosody, making voice conversion challenging. The study focused on bidirectional gender conversion while preserving speaker identity and linguistic content in the transformed speech. With the aid of pretrained speaker embeddings and pitch-extraction models, the system captured tonal and spectral characteristics specific to each language. Objective evaluation metrics, including F0, STOI, WER, and MCD, were calculated to quantify conversion accuracy and quality. The results indicate the potential of EGNVC for multilingual applications such as speech synthesis, dubbing, assistive communication, and personalized voice conversion.

Objective measures of speech quality and conversion accuracy were used to evaluate the EGNVC framework. Table 3 presents the results of Telugu voice-conversion experiments for both female-to-male (F→M) and male-to-female (M→F) transformations.

From Table 3, it is shown that the EGNVC framework successfully conducts speaker identity transformations with balanceable degrees between pitch adaptation, intelligibility and linguistic accuracy. The high values of F0 support the capability of the model to rescale pitch properly for performing gender conversion. The fact that WER is higher in female-to-male conversion means that it's more difficult to adapt lower pitched training to a male speaker then the other way around which had much better, i.e., lower, WER. It is possible that this so occurs as a result of the increased level of diversity in male voice characteristics, and an increase in the number of pitch movements involved from females into males. Furthermore, the larger MCD results from male to female indicate that converting lower-pitch voice into higher-pitch voice will cause significant spectral trans-formation and distortions. It is expected that the quality of converted speech in terms of intelligibility and naturalness will be better if we improve our spectral mapping process and prosody control. Gender conversion performance is validated on three Dravidian languages: Tamil, Malayalam and Kannada using Proposed EGNVC model.

In the female-to-male conversion reported in Table 3, the model achieved an F0 of 51.61 Hz, indicating adjustment of the pitch characteristics toward the lower vocal range of male speakers. The STOI score of 0.12 indicates low intelligibility of the transformed speech, potentially because of spectral alterations introduced during conversion. The WER of 4.24% further indicates distortion in the converted speech. The MCD score of 276.646 reflects the spectral difference between the source and target speech; a larger MCD indicates greater spectral modification. These findings suggest that, although the model can alter speaker identity, improved spectral alignment is required to achieve better naturalness.

For male-to-female conversion, the F0 was measured at 50.89 Hz. The STOI score remained 0.12, indicating the same intelligibility limitations observed in female-to-male conversion. However, the WER improved substantially to 0.98%, indicating that the converted speech preserved linguistic content more accurately than in the female-to-male conversion. The MCD value of 292.819 was slightly higher than that of the female-to-male conversion, reflecting the more extensive spectral changes required for this transformation.

Table 3 shows that the EGNVC framework performs speaker-identity transformation with varying trade-offs among pitch adaptation, intelligibility, and linguistic accuracy. The F0 values support the model's ability to rescale pitch for gender conversion. The higher WER in female-to-male conversion indicates that this direction may be more difficult than male-to-female conversion. The larger MCD observed for male-to-female conversion indicates that transforming a lower-pitched male voice into a higher-pitched female voice requires substantial spectral modification. Improved spectral mapping and prosody control are therefore needed to enhance the intelligibility and naturalness of the converted speech. The model's gender-conversion

performance was further evaluated in Tamil, Malayalam, and Kannada.

Tamil has a rich system of phonological contrasts, including long and short vowels, and therefore requires detailed spectral adjustment during voice conversion. For the female-to-male conversion reported in Table 4, the model reduced F0 to 26.92 Hz, indicating a shift toward the male pitch range. However, the STOI score of 0.08 indicates low intelligibility and suggests that more precise prosodic modification is required. The WER of 0.97% indicates good preservation of linguistic content, while the MCD score of 175.14 reflects a moderate degree of spectral modification. In the male-to-female conversion, F0 increased to 139.64 Hz, aligning with a higher-pitched female voice.

Intelligibility improved slightly, with a STOI score of 0.11, although the WER increased to 2.36%, indicating somewhat greater linguistic error. The MCD value of 177.41 was slightly higher than that of the female-to-male conversion, reflecting the greater difficulty of transforming a lower-pitched male voice into a female vocal register. Overall, the model achieved the intended pitch transformation, but the low STOI scores indicate persistent difficulty in maintaining speech clarity. Further improvements are required to increase intelligibility and reduce spectral distortion.

Malayalam has a complex phonemic inventory, including nasalized vowels, which makes voice conversion particularly challenging. In the female-to-male conversion reported in Table 5, the model reduced F0 to 25.41 Hz, indicating successful pitch transformation. However, intelligibility remained low (STOI = 0.08), suggesting that the transformation reduced speech clarity. The WER was 1.16%, indicating moderate preservation of linguistic content, while the MCD score of 156.06 reflected substantial spectral modification.

In male-to-female conversion, F0 increased substantially to 189.27 Hz, aligning with female vocal characteristics. The STOI value of 0.07 was slightly lower than that obtained for female-to-male conversion, indicating a small further reduction in clarity. The WER of 1.73% indicated additional errors in preserving linguistic information, while the MCD of 169.74 reflected the substantial spectral modification required for gender conversion.

Overall, the EGNVC model successfully modified pitch in both conversion directions. F0 decreased to 25.41 Hz in female-to-male conversion and increased to 189.27 Hz in male-to-female conversion, consistent with expected gender-related pitch differences. However, the STOI scores of 0.08 and 0.07 remained low, indicating that speech clarity requires improvement. The MCD values of 156.07 and 169.75 quantify the spectral changes required for speaker adaptation. Although the model effectively adjusts pitch and speaker identity, improved intelligibility and spectral consistency are needed to enhance overall voice quality.

Kannada, which contains a distinctive set of retroflex sounds and language-specific prosodic patterns, required detailed modification of both pitch and spectral content. Table 6 presents the performance of the proposed EGNVC model for Kannada female-to-male and male-to-female conversion.

Table 4. Performance Metrics of the Proposed EGNVC Model for Tamil Female-to-Male and Male-to-Female Conversion.

Metric	Value	
	Female to Male	Male to Female
F0	26.92	139.64
STOI	0.08	0.11
WER	0.97	2.36
MCD	175.14474487304688	177.4060516357422

Table 5. Performance Metrics of the Proposed EGNVC Model for Malayalam Female-to-Male and Male-to-Female Conversion

Metric	Value	
	Female to Male	Male to Female
F0	25.41	189.27
STO	0.08	0.07
WER	1.16	1.73
MCD	156.06671142578125	169.74652099609375

In female-to-male conversion, F0 was reduced to 23.58 Hz, shifting the speech toward the male pitch range. The STOI score of 0.09 indicated slightly better intelligibility than that obtained for Tamil and Malayalam. The WER of 1.58% suggested that most linguistic content was retained, with the remaining errors likely related to prosody and phoneme clarity. The MCD of 132.07 was the lowest among the three languages, indicating comparatively better spectral matching. For male-to-female conversion, F0 increased to 153.66 Hz, aligning with the target female vocal characteristics. Among the tested conversions, the STOI score of 0.12 was the highest, indicating comparatively better intelligibility. The WER of 1.32% indicated better linguistic preservation than in the female-to-male conversion. Spectral distortion nevertheless remained a concern, as the MCD value of 169.20 was comparable to those of other male-to-female transformations. These results demonstrate the successful application of EGNVC to Kannada voice conversion, with substantial pitch and spectral modification. F0 changed from 23.58 Hz in female-to-male conversion to 153.66 Hz in male-to-female conversion, confirming effective pitch adjustment. However, the STOI scores of 0.09 and 0.12 remained low, indicating that intelligibility can be further improved. The MCD values of 132.07 and 169.20 quantify the extent of spectral modification in the two conversion directions. Although the model preserves speaker characteristics and pitch reasonably well, further improvements in speech quality and naturalness are required for Kannada voice conversion.

4.1 Ablation Study

To quantify the contributions of the principal components of the proposed EGNVC framework, an ablation study was performed under identical training settings, including the same dataset partition, batch size, optimizer, learning-rate schedule, and number of epochs. Four languages were evaluated.

The ablation results of the proposed EGNVC framework are summarized in Table 7. The proposed EGNVC configuration combining Harvest and speaker embeddings achieved the best performance, indicating improved intelligibility and spectral quality. Removing Harvest reduced performance, showing that accurate pitch extraction contributes to improved prosody and naturalness.

Table 6. Performance Metrics of the Proposed EGNVC Model for Kannada Female-to-Male and Male-to-Female Conversion

Metric	Value	
	Female to Male	**Male to Female**
F0	23.58	153.66
STOI	0.09	0.12
WER	1.58	1.32
MCD	132.07159423828125	169.2061767578125

Table 7. Average Ablation Performance Across Four Indian Regional Languages

Model Variant	Avg STOI	Avg WER	Avg MCD
Without Harvest	0.09	2.08	205.31
Without Embedding	0.08	2.54	214.92
Proposed EGNVC (Harvest + Embedding)	0.10	1.79	193.64

Table 8. Comparative Performance Across Average Language Tasks

Method	STOI	WER	MCD	MOS	Speaker Similarity
AutoVC [29]	0.08	2.46	218.73	2.91 ± 0.18	3.02 ± 0.21
StarGANv2-VC [30]	0.09	2.21	206.55	3.18 ± 0.16	3.26 ± 0.19
RVC Baseline [32]	0.09	1.97	199.84	3.41 ± 0.14	3.48 ± 0.17
Proposed EGNVC	0.10	1.79	193.64	3.67 ± 0.13	3.74 ± 0.15

Removing speaker embeddings also reduced performance, confirming that embedding-guided conditioning is important for retaining target-speaker characteristics and achieving effective voice conversion. The proposed EGNVC framework was also compared with recent voice-conversion models using the same dataset split and preprocessing settings; the corresponding results are reported in Table 8. The EGNVC system showed reasonable performance for gender-dependent voice conversion in Tamil, Malayalam, and Kannada, although each language presented distinct challenges in pitch conversion, intelligibility, and spectral distortion.

Kannada conversions achieved comparatively better intelligibility, as indicated by higher STOI scores, whereas Malayalam conversions required greater spectral modification, as indicated by higher MCD values. WER was highest for Tamil male-to-female conversion, suggesting greater difficulty in preserving linguistic content in that direction. Although the model successfully altered pitch and speaker identity, spectral artifacts and intelligibility limitations remained, particularly in male-to-female conversion.

Future work should incorporate prosody-aware feature modelling and adaptive phoneme transformation to improve speech naturalness. Nevertheless, the model's performance in pitch transformation and linguistic preservation indicates potential for applications such as speech synthesis, dubbing, assistive technology, and multilingual voice adaptation.

5. Conclusion

In this study, an EGNVC framework was presented to improve the performance, generalization, and efficiency of voice conversion across multiple speakers and languages. The experimental results showed that the proposed method can improve pitch modification, linguistic preservation, and spectral correspondence. A key novelty of the approach is its ability to perform zero-shot voice conversion, enabling transformation of unseen speakers without training on a specific source-target pair. Unlike baseline systems that require large amounts of parallel data, the proposed EGNVC model generalized across different speaker identities and linguistic conditions, demonstrating potential for real-world applications. The Harvest pitch-extraction algorithm also enabled controlled pitch modification, contributing to the naturalness and expressiveness of the converted speech. The model was validated quantitatively. In female-to-male conversion, F0 values of 23.58 and 25.41 Hz were obtained for Kannada and Malayalam, respectively, whereas male-to-female conversion produced F0 values of 153.66 and 189.27 Hz. STOI scores ranged from 0.07 to 0.12, emphasizing the need for further improvement in speech intelligibility. Across all conversions, WER reached a maximum of 4.24%, while Kannada male-to-female conversion achieved a WER of 1.32%, indicating comparatively better linguistic preservation. MCD values ranged from 132.07 to 292.81, with higher values indicating more extensive spectral modification during voice adaptation.

References

[1] T. Walczyna, Z. Piotrowski, Overview of voice conversion methods based on deep learning. Applied sciences, 13(5), (2023) 3100. https://doi.org/10.3390/app13053100

[2] B. Li, J. Chen, Y. Xu, W. Li, Z. Liu, DRAW: Dual-Decoder-Based Robust Audio Watermarking Against Desynchronization and Replay Attacks. IEEE Transactions on Information Forensics and Security, 19, (2024) 6529 – 6544. https://doi.org/10.1109/TIFS.2024.3416047

[3] A. Triantafyllopoulos, B.W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E André, An overview of affective speech synthesis and conversion in the deep learning era. Proceedings of the IEEE, 111(10), (2023) 1355-1381. https://doi.org/10.1109/JPROC.2023.3250266

[4] Z. Yang, X. Jing, A. Triantafyllopoulos, M. Song, I. Aslan, B.W. Schuller, (2022) An overview & analysis of sequence-to-sequence emotional voice conversion. arXiv preprint arXiv:2203. 15873. https://doi.org/10.48550/arXiv.2203.15873

[5] W.C. Huan, L.P. Violeta, S. Liu, J. Shi, T. Toda, (2023) The singing voice conversion challenge 2023. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, Taipei, Taiwan. https://doi.org/10.1109/ASRU57964.2023.10389671

[6] M. I. Hussan, D. Saidulu, P.T. Anitha, A. Manikandan, P. Naresh, Object detection and recognition in real time using deep learning for visually impaired people. International Journal of Electrical and Electronics Research, 10(2), (2022) 80-86. https://doi.org/10.37391/ijeer.100205

[7] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, S. Poria, A review of deep learning techniques for speech processing. Information Fusion, 99, (2023) 101869. https://doi.org/10.1016/j.inffus.2023.101869

[8] J.H. Belz, L.M. Weilke, A. Winter, P. Hallgarten, E. Rukzio, T. Grosse-Puppendahl, Story-Driven: Exploring the Impact of Providing Real-time Context Information on Automated Storytelling. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology, 73, (2024) 1-15. https://doi.org/10.1145/3654777.3676372

[9] S. Aggarwal, S. Uttam, S. Garg, S. Garg, K. Jain, S. Aggarwal, Advancements in end-to-end audio style transformation: a differentiable approach for voice conversion and musical style transfer. AI, 6(1), (2025) 16. https://doi.org/10.3390/ai6010016

[10] C.W. Bang, C. Chun, Effective zero-shot multi-speaker text-to-speech technique using information perturbation and a speaker encoder. Sensors, 23(23), (2023) 9591. https://doi.org/10.3390/s23239591

[11] W. Slam, Y. Li, N. Urouvas, Frontier research on low-resource speech recognition technology. Sensors, 23(22), (2023) 9096. https://doi.org/10.3390/s23229096

[12] X. Xu, L. Shi, X. Chen, P. Lin, J. Lian, J. Chen, Z. Zhang, E.R. Hancock, Any-to-any voice conversion with multi-layer speaker adaptation and content supervision. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, (2023) 3431-3445. https://doi.org/10.1109/TASLP.2023.3306716

[13] A. Angra, H. Muralikrishna, D.A. Dinesh, V. Thenkanidiyoor, Exploring aggregated wav2vec 2.0 features and dual-stream TDNN for efficient spoken dialect identification. IEEE Access, 13, (2024) 3115-3129. https://doi.org/10.1109/ACCESS.2024.3523951

[14] K. Ezzine, J.D. Martino, M. Frikha, Any-to-one non-parallel voice conversion system using an autoregressive conversion model and lpcnet vocoder. Applied Sciences, 13(21), (2023) 11988. https://doi.org/10.3390/app132111988

[15] W.S. Hsu, G.T. Lin, W.H. Wang, Enhancing dysarthric voice conversion with fuzzy expectation maximization in diffusion models for phoneme prediction. Diagnostics, 14(23), (2024) 2693. https://doi.org/10.3390/diagnostics14232693

[16] D. Yook, G. Han, H.P. Chang, I.C. Yoo, Cyclediffusion: Voice conversion using cycle-consistent diffusion models. Applied Sciences, 14(20), (2024) 9595. https://doi.org/10.3390/app14209595

[17] K. Zhou, B. Sisman, R. Rana, B.W. Schuller, H. Li, Emotion intensity and its control for emotional voice conversion. IEEE Transactions on Affective Computing, 14(1), (2022) 31-48. https://doi.org/10.1109/TAFFC.2022.3175578

[18] D.S. Asudani, N.K. Nagwani, P. Singh, Impact of word embedding models on text analytics in deep learning environment: a review. Artificial intelligence review, 56(9). (2023): 10345-10425. https://doi.org/10.1007/s10462-023-10419-1

[19] M. Matassoni, S. Fong, A. Brutti, Speaker anonymization: Disentangling speaker features from pre-trained speech embeddings for voice conversion. Applied Sciences, 14(9), (2024) 3876. https://doi.org/10.3390/app14093876

[20] M. Amarjouf, E.H. Ibn Elhaj, M. Chami, K. Ezzine, J. Di Martino, Assessment of Self-Supervised Denoising Methods for Esophageal Speech Enhancement. Applied Sciences, 14(15), (2024) 6682. https://doi.org/10.3390/app14156682

[21] M. Zhang, Y. Zhou, Y. Ren, C. Zhang, X. Yin, H. Li, Refxvc: Cross-lingual voice conversion with enhanced reference leveraging. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, (2024) 4146-4156. https://doi.org/10.1109/TASLP.2024.3439996

[22] K. Guo, Y. Xu, S. Zhang. IAFF-VC: Any-to-Any Voice Conversion Using Attentional Feature Fusion. Circuits, Systems, and Signal Processing, 44 (2025) 8489-8509. https://doi.org/10.1007/s00034-025-03198-3

[23] R. Yamamoto, R. Yoneyama, L.P. Violeta, W.C. Huang, T. Toda, (2023) A comparative study of voice conversion models with large-scale speech and singing data: The T13 systems for the singing voice conversion challenge 2023. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, Taipei, Taiwan. https://doi.org/10.1109/ASRU57964.2023.10389779

[24] B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, (2020) 132-157. https://doi.org/10.1109/TASLP.2020.3038524

[25] S. Chen, Y. Gu, J. Cui, J. Zhang, R. Chen, L. Dai, Lcm-svc: Latent diffusion model based singing voice conversion with inference acceleration via latent consistency distillation. In 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, Beijing, China. https://doi.org/10.1109/ISCSLP63861.2024.10800358

[26] W. Lu, X. Zhao, N. Guo, Y. Li, J. Wei, J. Tao, J. Dang, One-shot emotional voice conversion based on feature separation. Speech Communication, 143, (2022) 1-9. https://doi.org/10.1016/j.specom.2022.07.001

[27] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, S. Seki, (2020). Voicegrad: Non-parallel any-to-many voice conversion with annealed langevin dynamics. Arxiv Preprint Arxiv: 2010.02977. https://doi.org/10.48550/arXiv.2010.02977

[28] H. Guo, C. Liu, C.T. Ishi, H. Ishiguro, (2023) Using joint training speaker encoder with consistency loss to achieve cross-lingual voice conversion and expressive voice conversion. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) IEEE, Taipei, Taiwan.

https://doi.org/10.1109/ASRU57964.2023.10389651

[29] K. Qian, Y. Zhang, S. Chang, X. Yang, M. Hasegawa-Johnson, AutoVC: Zero-shot voice style transfer with only autoencoder loss. In Proceedings of the 36th International Conference on Machine Learning, PMLR, 97, (2019) 5210–5219.

[30] Y.A. Li, A. Zare, N, Mesgarani, (2021) StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. Arxiv Preprint Arxiv: 2107.10394. https://doi.org/10.48550/arXiv.2107.10394

[31] A. Das, S. Ghosh, T. Polzehl, I. Siegert, S. Stober, (2023). StarGAN-VC++: Towards emotion preserving voice conversion using deep embeddings. Arxiv Preprint Arxiv: 2309.07592. https://doi.org/10.48550/arXiv.2309.07592

[32] RVC-Project, (2023). Retrieval-based-Voice-Conversion-WebUI. GitHub repository. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI

Authors Contribution Statement

A. Bala Raju: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Validation, Visualization, Writing – Original Draft. S. P. Singh: Supervision, Methodology, Validation, Resources, Writing – Review & Editing. Dhiraj Sunehra: Supervision, Investigation, Validation, Project administration, Writing – Review & Editing. All authors contributed to drafting and revising the manuscript.

Funding

The authors declare that this study received no specific funding

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License

© The Author(s) 2026. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License.