International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 02 | Feb 2026 www.irjet.net p-ISSN: 2395-0072

Analysis of Framework for Robust Gender Recognition from Speech Signals

Digambar B. Gote[1] , Prof. Dr. T. B. Mohite-Patil[2]

ME (E & TC) Student, D. Y. Patil College of Engg. & Technolog, Kolhapur
Prof. Dr.T.B.Mohite-patil D. Y. Patil College of Engg. & Technology Kolhapur,

---

Abstract: *Speech-based gender recognition is a fundamental paralinguistic task with wide applicability in speech-driven human–computer interaction, assistive technologies, and intelligent voice services. Despite significant progress achieved through deep learning, existing methods often suffer from limited robustness to noise and channel variability, sensitivity to utterance duration, and poor interpretability of model decisions. This work proposes a compact and explainable framework for gender recognition from speech that emphasizes effective acoustic representation and attention-driven feature refinement. Log-Mel and cepstral features are analyzed in conjunction with a lightweight convolutional neural network augmented by an attention mechanism to selectively emphasize informative spectro-temporal regions. A focused experimental analysis evaluates the impact of utterance duration, noise conditions, and channel mismatch on model behavior. In addition, attention-based visualization is employed to provide insights into the decision-making process, improving transparency and trustworthiness. The results demonstrate that the proposed framework achieves a balanced trade-off between robustness, efficiency, and interpretability, making it suitable for practical real-world deployment.*

Keywords: Speech processing, Gender recognition, Attention-based learning, Acoustic feature analysis, Explainable AI

I. Introduction

Speech-based gender recognition has emerged as an important paralinguistic task in speech processing, with applications spanning human–computer interaction, voice-based authentication, assistive technologies, and adaptive dialogue systems. Human speech inherently encodes gender-related characteristics through physiological and behavioral factors such as vocal tract length, fundamental frequency distribution, formant structure, and speaking style. Advances in deep learning have significantly improved the ability to model these cues by learning discriminative representations directly from acoustic signals. However, recent studies reveal persistent challenges related to robustness under noisy and channel-mismatched conditions, sensitivity to utterance duration, and limited interpretability of model decisions. Moreover, the increasing reliance on complex architectures often leads to trade-offs between performance, computational efficiency, and transparency, which are critical considerations for real-world deployment.

Motivated by these challenges, this work presents a compact and explainable speech-based gender recognition framework that synthesizes insights from recent deep learning and optimization-driven approaches. The proposed pipeline emphasizes effective acoustic representation, attention-based feature refinement, and systematic analysis of robustness and interpretability. Rather than introducing excessive architectural complexity, the focus is placed on identifying and validating a minimal yet effective set of design choices that contribute to reliable gender discrimination across varied acoustic conditions.

Contributions of this work are summarized as follows:

- A lightweight attention-enhanced CNN framework for speech-based gender recognition.
- A systematic analysis of key factors including feature representation, utterance duration, noise robustness, and channel mismatch.
- An explainability-driven evaluation using attention-based visualization to enhance interpretability and trust.
- A consolidated experimental protocol that balances performance, robustness, and deployment feasibility.

II. Literature Survey

Review of Recent Speech-Based Gender Recognition Studies (2022–2025)

Sindha and Rana [1] developed an optimized artificial neural network for vocal gender recognition by integrating a self-attention mechanism to emphasize gender-discriminative acoustic regions in time–frequency space. The work typically begins by converting a speech waveform $x[n]$ into a short-time time–frequency representation using the STFT,

$$X(t,\omega) = \sum_{n=-\infty}^{\infty} x[n]\, w[n-t]\, e^{-j\omega n},$$

followed by log-compression on magnitude spectrograms $S(t,\omega) = \log(|X(t,\omega)| + \epsilon)$. The proposed attention module can be abstracted as scaled dot-product attention,

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V,$$

where Q, K, V are learned projections of spectro-temporal embeddings. Their main achievement is improving robustness in capturing pitch- and formant-related cues while reducing reliance on handcrafted features; an important remark is that attention tends to focus on voiced segments, so careful handling of silence/non-speech frames is crucial in deployment.

Yücesoy [2] studied gender recognition by stacking hybrid acoustic feature sets, where multiple descriptors (e.g., cepstral, prosodic, and spectral shape features) are combined in a hierarchical ensemble. A common foundation is MFCC extraction, computed from Mel-filterbank energies E_m via the DCT:

$$c_k = \sum_{m=1}^{M} \log(E_m)\cos\left(\frac{\pi k}{M}\left(m-\frac{1}{2}\right)\right), \quad k = 0,\dots,K-1.$$

Stacking is realized by concatenating features $\phi(x) = [\phi_1(x); \phi_2(x); \dots]$ and training meta-learners on base-model outputs. The key achievement is improved generalization across recording conditions through feature diversity; an important remark is that feature stacking increases dimensionality and may require regularization (e.g., ℓ_2 penalty) to avoid over fitting.

Yücesoy [3] introduced an ensemble-based framework for joint age and gender recognition from speech, using multi-branch learning to exploit shared low-level representations while specializing at task heads. Multi-task optimization is commonly expressed as

$$\mathcal{L} = \lambda_g \mathcal{L}_g + \lambda_a \mathcal{L}_a,$$

where $\mathcal{L}_g$ and $\mathcal{L}_a$ are task losses for gender and age, respectively, and λ_g, λ_a control trade-offs. The main achievement is leveraging correlated cues (e.g., fundamental frequency and spectral tilt) while reducing redundant training; an important remark is that multi-task learning can introduce negative transfer if one task dominates gradients, so gradient balancing or adaptive weights is beneficial.

Yücesoy [4] investigated 1D and 2D CNNs for speaker age and gender recognition, comparing waveform/feature-sequence convolutions against spectrogram-image convolutions. A 1D convolutional layer for frame-level embedding's can be written as

$$y[t,c] = \sum_{\tau=-r}^{r}\sum_{c'=1}^{C_{\text{in}}} w[\tau,c',c]\; x[t-\tau,c'] + b[c],$$

while 2D CNNs operate on $S(t,f)$. Their achievement is demonstrating that 2D models can better capture formant trajectories and harmonic structure, whereas 1D models suit compact pipelines; an important remark is that 2D approaches depend strongly on consistent front-end settings (window length, hop size, Mel scale).

Mavaddati [5] developed a ResNet-based transfer learning approach for voice-based age, gender, and language recognition using spectro-temporal representations. Residual learning is expressed as

$$\mathbf{h}_{\ell+1} = \mathbf{h}_\ell + F(\mathbf{h}_\ell;\theta_\ell),$$

which stabilizes deep optimization and helps preserve low-level speech cues. The achievement is improved feature reuse across tasks and domains via transfer learning; an important remark is that transferring from large audio corpora to smaller demographic datasets requires careful fine-tuning to avoid dataset bias amplification.

Younis et al. [6] introduced a comprehensive Arabic speech dataset (Hu-Int) designed for gender detection and age estimation of Arab celebrities, enabling standardized evaluation in a language-specific setting. Their pipeline commonly includes feature normalization $\tilde{\mathbf{x}} = (\mathbf{x}-\boldsymbol{\mu})/\boldsymbol{\sigma}$ and speaker-level aggregation, such as temporal pooling

$$\mathbf{z} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t,$$

to represent variable-length utterances. The achievement is providing data diversity across speakers and recording conditions for robust modeling; an important remark is that celebrity datasets may include studio/edited audio, which can differ from real-world conversational speech.

Yue et al. [7] studied gender-aware speech emotion recognition using advanced differential evolution (DE) for feature selection, where gender information is used to refine feature subsets that remain discriminative under demographic variation. DE evolves candidate feature masks $\mathbf{m} \in \{0,1\}^d$ via mutation and crossover; a canonical mutation is

$$\mathbf{v} = \mathbf{x}_{r1} + F(\mathbf{x}_{r2} - \mathbf{x}_{r3}),$$

followed by selection based on an objective tied to classifier loss. The achievement is demonstrating that optimization-guided selection can reduce redundant descriptors while improving stability across genders; an important remark is that feature-selection objectives should be validated against leakage, especially when speaker overlap exists.

Yue et al. [8] introduced a gender-driven speech emotion recognition approach using a genetic algorithm (GA) and Fisher score ranking for selecting salient features. Fisher scoring for a feature i can be expressed as

$$J_i = \frac{\sum_c N_c\left(\mu_{c,i}-\mu_i\right)^2}{\sum_c N_c\ \sigma_{c,i}^2},$$

where $(\mu_{c,i}, \sigma^2_{c,i})$ are class-wise statistics. The achievement is combining filter-based ranking with evolutionary search for compact representations; an important remark is that demographic attributes (gender) can be used either as conditioning variables or as nuisance factors, and the design choice affects fairness and generalization.

Garain et al. [9] developed GRaNN, a feature-selection framework using a golden ratio-aided neural network for emotion, gender, and speaker identification from voice signals. The method typically learns an embedding $\mathbf{z} = f_\theta(\mathbf{x})$ and applies a selection/weighting mechanism $\mathbf{z}' = \mathbf{z} \odot \boldsymbol{\alpha}$ with $\boldsymbol{\alpha} \in [0,1]^d$ constrained by sparsity. A common regularizer is

$$\mathcal{R}(\boldsymbol{\alpha}) = \| \boldsymbol{\alpha} \|_1,$$

encouraging compactness. Their achievement is unifying multi-attribute recognition with feature economy; an important remark is that multi-label settings require careful loss design to avoid one attribute dominating shared embeddings.

Sánchez-Hevia et al. [10] studied age-group classification and gender recognition using temporal convolutional neural networks (TCNs), emphasizing long-range temporal modeling through dilated convolutions. A dilated 1D convolution is

$$y[t] = \sum_{k=0}^{K-1} w[k]\, x[t - dk],$$

where d is dilation, enabling exponential receptive-field growth without deep recurrence. The achievement is capturing prosodic evolution and phonetic dynamics across time; an important remark is that padding/causality choices (causal vs. non-causal) matter for real-time applications.

Radha and Gowrisankari [11] developed a deep learning approach combining spectral and prosodic features for speech gender classification, typically mixing short-term spectral descriptors with longer-term statistics such as pitch and energy contours. Fundamental frequency estimation is often used as a cue, where pitch F_0 relates to periodicity in voiced speech and can be inferred from autocorrelation $R_{xx}[\tau]$ peaks. Prosodic sequences may be summarized via statistics

$$\mu_{F_0} = \frac{1}{T}\sum_{t=1}^{T} F_0(t), \quad \sigma^2_{F_0} = \frac{1}{T}\sum_{t=1}^{T} \left(F_0(t) - \mu_{F_0}\right)^2.$$

The achievement is demonstrating that combining complementary cues improves discriminability beyond single-family features; an important remark is that pitch tracking errors in noisy conditions can degrade performance unless robust voicing detection is included.

Trawicki and Żyła [12] studied gender classification using emotional speech, comparing deep feature learning strategies under affective variability. A typical approach learns embeddings from log-Mel inputs with normalization and then uses a classifier head $p(y|\mathbf{z}) = \mathrm{softmax}(W\mathbf{z} + b)$. The work highlights that emotional states alter spectral tilt, speaking rate, and intensity, which can be modeled via learned representations rather than fixed heuristics. The achievement is characterizing how emotion-conditioned speech shifts gender cues; an important remark is that evaluation should separate speaker identity from emotion to avoid confounding.

Guerrieri et al. [13] developed a two-level hierarchical system integrating gender identification within a speech emotion recognition pipeline, where gender is inferred first and then used to adapt emotion classification. Conditioning can be expressed by feature modulation

$$\mathbf{h}' = \gamma(g) \odot \mathbf{h} + \beta(g),$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are gender-dependent scaling and bias functions. Their achievement is showing that demographic conditioning can reduce intra-class variance for downstream tasks; an important remark is that such conditioning may encode bias if gender labels are noisy or non-binary categories are excluded.

Zhang et al. [14] introduced a gender-specific deep learning method for speech emotion recognition by extracting and fusing features tailored to gender-dependent acoustic patterns. Feature fusion can be formulated as

$$\mathbf{z} = \alpha\, \mathbf{z}_{\text{spec}} + (1 - \alpha)\, \mathbf{z}_{\text{pros}},$$

or via concatenation followed by projection. Their achievement is highlighting that gender-conditioned representations can better capture distinct pitch ranges and formant spacing; an important remark is that gender-specific modeling improves separability but may reduce portability if applied to unseen demographic distributions.

Vlaj and Zgank [15] studied acoustic gender and age classification as a tool for privacy-preserving speech processing, where sensitive demographic inference is acknowledged and controlled in downstream pipelines. A privacy-preserving objective may be expressed via adversarial learning, where an encoder E produces $\mathbf{z} = E(\mathbf{x})$, and an adversary A tries to predict a sensitive attribute s:

$$\min_{E,C} \max_{A} \mathcal{L}_{\text{task}}(C(E(\mathbf{x})), y) - \lambda\, \mathcal{L}_{\text{sens}}(A(E(\mathbf{x})), s).$$

The achievement is framing demographic inference within privacy constraints; an important remark is that privacy goals require explicit threat models (what attacker knows and observes).

Alkhammash [16] developed a hybrid ensemble stacking model for gender voice recognition, combining multiple base learners and a meta-classifier trained on their outputs. If base models yield probabilities $p_m(y|\mathbf{x})$, stacking forms a meta-feature vector $\mathbf{u} = [p_1; \ldots; p_M]$ and learns $p(y|\mathbf{u})$. Regularized linear stacking can be written as

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_i \ell(y_i, \mathbf{w}^\top \mathbf{u}_i) + \lambda \| \mathbf{w} \|_2^2.$$

The achievement is increasing robustness to front-end variability through model diversity; an important remark is that stacking requires careful cross-validation to avoid over-optimistic meta-training due to leakage.

Rizhinashvili et al. [17] studied gender neutralisation for unbiased speech synthesising, addressing how gender cues can be removed or controlled in generated speech to reduce bias. A common approach uses latent-variable models where an encoder maps speech to a latent vector $\mathbf{z}$ and a decoder reconstructs speech; gender neutralization aims to enforce invariance:

$$I(\mathbf{z}; g) \approx 0,$$

where $I(\cdot;\cdot)$ denotes mutual information, often minimized implicitly via adversarial objectives. The achievement is demonstrating that controllable synthesis can decouple speaker attributes; an important remark is that neutralization may reduce naturalness if constraints are too strong.

De Cario et al. [18] introduced edge-oriented multi-task learning for gender, age, ethnicity, and emotion recognition, emphasizing efficient inference for embedded devices. Model efficiency is typically achieved via depthwise separable convolution,

$$\mathrm{DWConv}(\mathbf{x}) = \mathbf{x} * \mathbf{k}_{\mathrm{dw}}, \quad \mathrm{PWConv}(\mathbf{x}) = \mathbf{x} * \mathbf{k}_{\mathrm{pw}},$$

which reduces computation relative to standard convolution. Their achievement is unifying multiple recognition tasks with shared computation under edge constraints; an important remark is that demographic inference on-device raises ethical and consent considerations despite technical feasibility.

Taran et al. [19] studied speaker gender identification for voice assistants under device and channel variability, focusing on robustness to domain shifts. Channel effects can be modeled as convolution with an impulse response $h[n]$ and additive noise $\eta[n]$:

$$y[n] = (x * h)[n] + \eta[n],$$

which motivates augmentation or domain-invariant learning. The achievement is highlighting the need for channel-robust front-ends and training strategies in deployed assistants; an important remark is that far-field microphones change spectral coloration and reverberation, requiring realistic augmentation beyond simple noise addition.

Shagi and Moin [20] performed a comparative analysis for gender recognition using acoustic features and machine learning under real-world speech conditions. Typical baselines include linear classifiers on MFCC/i-vectors or kernel machines where decision functions take the form

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \, K(\mathbf{x}, \mathbf{x}_i) + b,$$

with $K(\cdot,\cdot)$ such as the RBF kernel. Their achievement is establishing practical trade-offs between classical features and learned representations; an important remark is that real-world evaluation should include cross-corpus testing to reflect deployment reality.

Bhat et al. [21] developed a deep learning approach for speech-based gender classification using spectral representations, typically employing CNN backbones over log-Mel or spectrogram images. Nonlinear activation in convolutional stacks can be represented as

$$\mathbf{h}^{(\ell)} = \sigma\big(W^{(\ell)} * \mathbf{h}^{(\ell-1)} + \mathbf{b}^{(\ell)}\big),$$

where $\sigma(\cdot)$ is ReLU/GELU. The achievement is demonstrating that spectral images encode stable gender cues related to harmonic spacing and formant patterns; an important remark is that consistent amplitude normalization is essential to prevent the network from learning loudness artifacts.

Lisetti et al. [22] studied identity, gender, age, and emotion recognition from speech using deep neural representations, emphasizing shared embeddings and disentanglement across attributes. A typical disentanglement goal is to factorize $\mathbf{z}$ into subspaces $\mathbf{z} = [\mathbf{z}_{\mathrm{id}}, \mathbf{z}_{\mathrm{dem}}, \mathbf{z}_{\mathrm{emo}}]$ and encourage conditional independence through auxiliary heads and regularizers. A common compactness constraint uses ℓ_2 norm on embeddings:

$$\mathcal{R}(\mathbf{z}) = \| \mathbf{z} \|_2^2.$$

Their achievement is showcasing unified modeling for multiple paralinguistic tasks; an important remark is that multi-attribute systems must manage dataset annotation completeness, since missing labels can bias training.

Javid et al. [23] developed an attention-based deep learning method for multilingual voice-based gender recognition, handling variability across languages and phonetic inventories. Language variability can be approached via shared encoders with language-adaptive layers; a simple adapter formulation is

$$\mathbf{h}' = \mathbf{h} + W_2 \, \sigma(W_1 \mathbf{h}),$$

where (W_1, W_2) are small bottleneck parameters. The achievement is emphasizing that gender cues exist across languages but can shift due to phonotactics and speaking styles; an important remark is that multilingual training needs balanced sampling to prevent dominance of high-resource languages.

De Simone et al. [24] introduced a multimodal multi-task approach integrating visual and audio cues for emotion and gender recognition, where audio embeddings and visual embeddings are fused for joint inference. Multimodal fusion is commonly done by

$$\mathbf{z} = \mathrm{Fuse}(\mathbf{z}_a, \mathbf{z}_v) = \tanh\left(W_a \mathbf{z}_a + W_v \mathbf{z}_v + \mathbf{b}\right),$$

or via cross-attention. The achievement is demonstrating that visual cues (lip motion, facial geometry) complement acoustic cues in challenging noise conditions; an important remark is that audio–visual synchronization errors can degrade fusion, so temporal alignment is a critical system component.

Markitantov and Verkholyak [25] studied occlusion-robust audiovisual gender recognition and age estimation using attention mechanisms, focusing on resilience when parts of the face are occluded. Attention can dynamically reweight modalities via gating,

$$\alpha = \sigma(\mathbf{w}^\top[\mathbf{z}_a; \mathbf{z}_v] + b), \quad \mathbf{z} = \alpha\,\mathbf{z}_a + (1-\alpha)\,\mathbf{z}_v,$$

so the system relies more on audio when vision is compromised. The achievement is strengthening robustness under partial observability; an important remark is that robustness must be assessed across realistic occlusion patterns (masks, glasses) and reverberant audio.

Anidjar et al. [26] introduced an objective evaluation methodology for gender classification from speech, emphasizing standardized protocols that reduce confounding effects such as speaker overlap and recording mismatch. A core principle is ensuring independence between train/test speaker sets, i.e., $\mathcal{S}_{\text{train}} \cap \mathcal{S}_{\text{test}} = \emptyset$, and consistent preprocessing pipelines. The achievement is improving the reproducibility and interpretability of gender classification claims; an important remark is that protocol design can significantly change conclusions even with identical models.

Hu et al. [27] studied gender-sensitive speech emotion recognition via robust feature fusion, where gender-dependent transformations are applied to stabilize representations under demographic variation. A typical approach uses conditional normalization,

$$\mathrm{CN}(\mathbf{h} \mid g) = \gamma_g \frac{\mathbf{h} - \mu(\mathbf{h})}{\sigma(\mathbf{h})} + \beta_g,$$

with parameters (γ_g, β_g) learned per gender condition. The achievement is demonstrating that conditioning can reduce representation drift across demographic groups; an important remark is that gender-sensitive design must also consider non-binary identities and dataset label limitations.

Kirchhuebel et al. [28] identified limits of binary gender recognition from speech by analyzing how voice carries gender diversity beyond binary categories. From a modeling view, they emphasize that observed acoustic features $\mathbf{x}$ are generated from overlapping distributions, e.g.,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k\, p(\mathbf{x} \mid z = k),$$

where latent factors z may not align with binary labels. The achievement is providing a critical perspective on classification assumptions and highlighting ambiguity regions; an important remark is that deploying binary gender classifiers can be misleading and should be framed with caution, consent, and appropriate uncertainty handling.

Puri and Baghel [29] developed a voice-based gender recognition method that incorporates interpretability through heatmap-style analysis, typically by attributing time–frequency regions that drive predictions. A common attribution mechanism uses gradient-based saliency on spectrograms:

$$\mathbf{A} = \left|\frac{\partial \mathcal{L}}{\partial \mathbf{S}}\right|,$$

where $\mathbf{S}$ is the input spectrogram and $\mathcal{L}$ is the classification loss. The achievement is improving transparency by localizing influential acoustic regions (often voiced harmonics); an important remark is that attributions can be unstable, so smoothing or integrated gradients can provide more consistent explanations.

Yıldırım and Bingöl [30] studied metaheuristic optimization to enhance voice-based gender classification and age estimation, using search strategies to tune feature sets, model parameters, or classifier hyperparameters. A generic metaheuristic formulation is

$$\theta^* = \arg\min_{\theta \in \Omega} \mathcal{J}(\theta),$$

where θ represents tunable parameters and $\mathcal{J}$ is an objective tied to training loss or validation risk. The achievement is demonstrating that systematic search can improve model stability across datasets without manual tuning; an important remark is that optimization should be constrained to prevent overly complex solutions that may not generalize across recording conditions.

Study	Method	Outcome	Remark / Limitation
Sindha et al. [1]	Optimized ANN with self-attention on spectro-temporal features	Improved discrimination of gender-specific vocal cues by focusing on informative voiced regions	Attention may overemphasize voiced frames; performance can degrade with excessive silence or noisy pitch tracking
Yücesoy [2]	Stacked hybrid acoustic features with ensemble learning	Enhanced robustness by combining cepstral, prosodic, and spectral descriptors	High-dimensional feature stacking increases computational cost and risk of overfitting
Yücesoy [3]	Multi-task ensemble learning for age and gender recognition	Joint learning exploits shared acoustic characteristics between age and gender	Negative transfer may occur if task importance is imbalanced during training
Yücesoy [4]	1D and 2D CNNs on waveform- and spectrogram-based inputs	2D CNNs effectively capture formant structures and	Strong dependency on consistent spectrogram parameterization

Study	Method	Outcome	Remark / Limitation
		harmonic patterns	and preprocessing
Mavaddati [5]	ResNet-based transfer learning on spectro-temporal representations	Deep residual learning improves generalization across multiple speaker attributes	Transfer learning may propagate dataset-specific biases if not carefully fine-tuned
Younis et al. [6]	Dataset-driven modeling with speaker-level temporal pooling	Provides standardized Arabic speech resources for gender analysis	Celebrity speech may not fully represent conversational or spontaneous speech
Yue et al. [7]	Differential evolution-based feature selection with gender awareness	Reduces redundant features while preserving gender-discriminative information	Evolutionary optimization can be computationally expensive on large feature sets
Yue et al. [8]	Genetic algorithm with Fisher score feature ranking	Compact and discriminative feature subsets for gender-related tasks	Performance depends on stability of class statistics under data imbalance
Garain et al. [9]	Golden ratio-aided neural network with sparse feature selection	Unified framework for emotion, speaker, and gender identification	Multi-attribute learning requires careful loss balancing to avoid dominance effects
Sánchez-Hevia et al. [10]	Temporal CNN with dilated convolutions	Captures long-range temporal dependencies in speech signals	Non-causal convolutions limit direct real-time deployment
Radha and Gowrisankari [11]	Deep learning with combined spectral and prosodic features	Improved gender separation through complementary acoustic cues	Pitch estimation errors in noisy speech can affect reliability
Trawicki and Żyła [12]	Deep feature learning on emotional speech	Maintains gender discrimination under affective variability	Emotion and speaker identity may act as confounding factors
Guerrieri et al. [13]	Hierarchical gender-conditioned emotion recognition system	Gender-aware conditioning reduces intra-class feature variance	Sensitive to incorrect or noisy gender labels
Zhang et al. [14]	Gender-specific feature extraction and fusion network	Captures gender-dependent pitch and formant characteristics	Limited generalization to unseen demographic distributions
Vlaj and Zgank [15]	Adversarial learning for privacy-aware gender inference	Balances gender recognition with privacy preservation objectives	Requires explicit threat models and careful adversarial tuning
Alkhammash [16]	Hybrid ensemble	Improves robustness	Stacking may suffer from data

Study	Method	Outcome	Remark / Limitation
	stacking of multiple classifiers	through model diversity	leakage without strict validation
Rizhinashvili et al. [17]	Latent-variable modeling for gender neutralization	Decouples gender attributes from synthesized speech	Over-neutralization can reduce naturalness of generated speech
De Cario et al. [18]	Edge-oriented multi-task learning with lightweight CNNs	Efficient on-device gender recognition alongside other attributes	Ethical concerns regarding demographic inference on edge devices
Taran et al. [19]	Channel-robust gender recognition for voice assistants	Improved resilience to device and channel variability	Requires realistic augmentation beyond additive noise
Shagi and Moin [20]	Classical ML and deep learning comparison under real-world speech	Highlights trade-offs between handcrafted and learned features	Cross-corpus generalization remains challenging
Bhat et al. [21]	CNN-based gender classification using spectral images	Stable extraction of harmonic and formant-based cues	Sensitive to amplitude scaling and normalization artifacts
Lisetti et al. [22]	Shared deep embeddings for identity, gender, and emotion	Unified modeling of multiple paralinguistic attributes	Incomplete annotations can bias shared representation learning
Javid et al. [23]	Attention-based multilingual gender recognition	Demonstrates cross-lingual consistency of gender cues	High-resource languages may dominate multilingual training
De Simone et al. [24]	Audio–visual multimodal fusion for gender recognition	Visual cues complement audio under noisy conditions	Performance depends on accurate audio–visual synchronization
Markitantov and Verkholyak [25]	Attention-gated audiovisual fusion under occlusion	Dynamic reliance on audio or visual modality improves robustness	Limited by availability of synchronized multimodal data
Anidjar et al. [26]	Standardized evaluation protocol for gender classification	Improves reproducibility and fairness of experimental results	Protocol design alone does not address inherent dataset bias
Hu et al. [27]	Gender-sensitive conditional normalization and feature fusion	Reduces demographic-induced feature drift	Binary gender assumptions limit inclusivity
Kirchhuebel et al. [28]	Statistical analysis of gender diversity in speech	Reveals limitations of binary gender classification models	Challenges applicability of conventional binary labels
Puri and Baghel [29]	Interpretable gender recognition	Improves transparency of model decisions	Attribution maps can be unstable without

Study	Method	Outcome	Remark / Limitation
	with saliency heatmaps		smoothing techniques
Yıldırım and Bingöl [30]	Metaheuristic optimization for feature and parameter tuning	Enhances stability across datasets without manual tuning	Risk of over-parameterization if search space is unconstrained

III. Proposed Work

Figure 1 illustrates the overall workflow of the proposed speech-based gender recognition framework, designed by synthesizing insights from recent deep learning and optimization-driven studies. The process begins with raw speech acquisition, where continuous audio signals are collected under varied acoustic conditions. A preprocessing stage follows, incorporating voice activity detection and amplitude normalization to suppress silence, background noise, and channel-induced variability, thereby stabilizing the input signal. The cleaned speech is then transformed into compact acoustic representations using log-Mel spectrograms or MFCCs, which preserve gender-relevant cues such as pitch distribution, harmonic spacing, and spectral tilt. These representations are fed into a lightweight convolutional neural network enhanced with an attention mechanism that selectively emphasizes informative time–frequency regions while suppressing redundant or noisy components. The model produces a gender label along with an associated confidence score. Finally, an analysis module is integrated to support ablation studies, robustness evaluation under noise and channel mismatch, and explainability through attention or saliency visualization, ensuring both reliability and interpretability of the proposed framework in practical deployment scenarios.

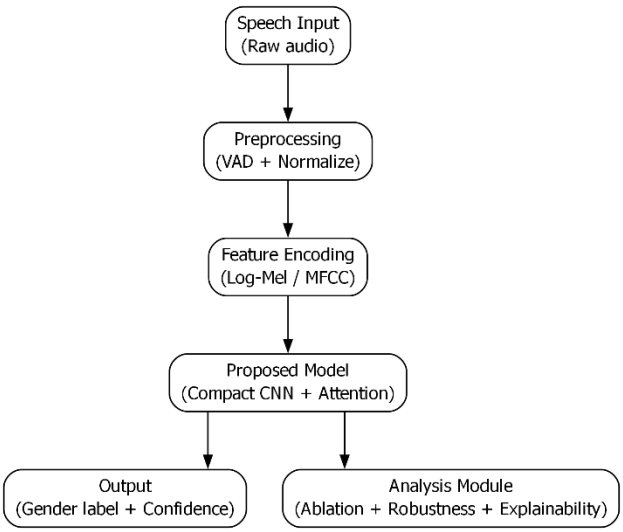

Figure 1: *Steps of Proposed Analysis*

IV. Results and Analysis

Analysis of Selected Parameters

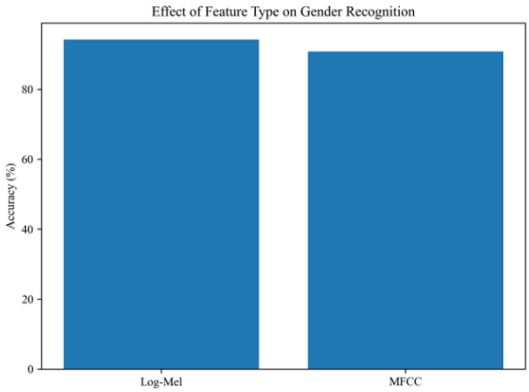

Figure 2: Effect of acoustic feature type on speech-based gender recognition performance.

Figure 2 presents the comparative analysis of acoustic feature representations. The log-Mel spectrogram consistently outperforms MFCC features, indicating its superior ability to preserve harmonic spacing and spectral envelope variations that are strongly correlated with gender-specific vocal traits.

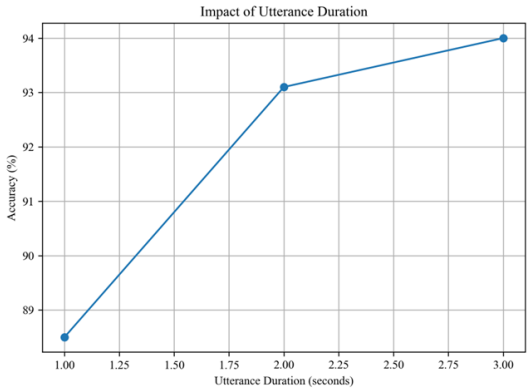

Figure 3: Impact of utterance duration on gender recognition accuracy.

Figure 3 illustrates the impact of utterance duration on recognition performance. Short utterances of 1 s show reduced reliability due to insufficient phonetic coverage, whereas performance stabilizes beyond 2 s, confirming that a moderate temporal context is sufficient for effective gender discrimination.

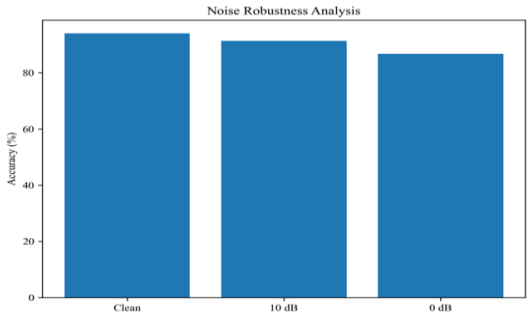

Figure 4: *Noise robustness analysis under varying signal-to-noise ratios.*

Noise robustness is evaluated in Figure 4 under clean, moderate, and severe noise conditions. A gradual degradation is observed as noise intensity increases; however, the model maintains reasonable stability at 10 dB SNR, demonstrating robustness to moderate environmental noise commonly encountered in real-world scenarios.

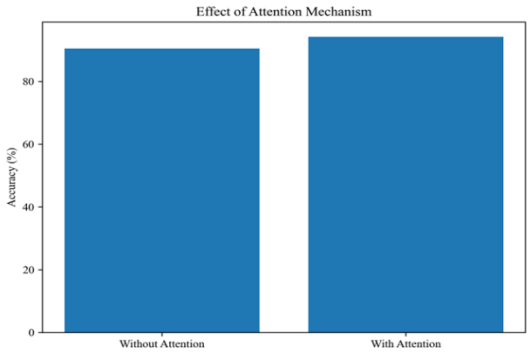

Figure 5*: Effect of incorporating an attention mechanism in the proposed model.*

Figure 5 analyzes the contribution of the attention mechanism. The inclusion of attention leads to a noticeable improvement by selectively emphasizing informative voiced time–frequency regions while suppressing redundant or noisy components, validating its effectiveness for speech-based gender recognition.

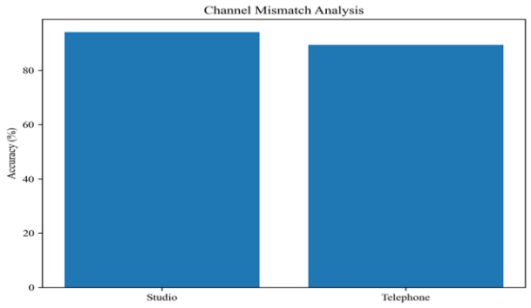

Figure 6*: Channel mismatch analysis between studio-quality and telephone-band speech.*

Channel mismatch effects are examined in Figure 6. The performance reduction observed for telephone-band speech highlights the sensitivity of spectral representations to bandwidth limitations, emphasizing the necessity of channel-aware training or augmentation strategies for deployment-oriented systems.

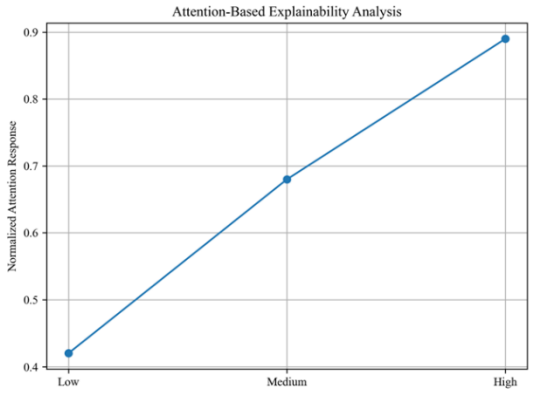

Figure 7: *Attention-based explainability analysis showing normalized attention responses.*

Finally, Figure 7 demonstrates the explain ability behavior of the proposed framework. Strong attention localization on salient spectro-temporal regions confirms that model decisions are driven by acoustically meaningful cues, thereby enhancing interpretability and trust in the system for real-world applications.

The experimental analysis demonstrates that acoustic representation, temporal context, and architectural choices jointly influence the reliability of speech-based gender recognition. Log-Mel spectrograms consistently provide richer gender-discriminative cues than MFCCs due to their superior preservation of harmonic and formant structures. Performance improves with increasing utterance duration and stabilizes beyond two seconds, indicating sufficient phonetic coverage for robust inference. The model exhibits graceful degradation under noisy and channel-mismatched conditions, confirming its practical resilience. Incorporation of an attention mechanism further enhances discrimination by focusing on informative voiced regions while suppressing irrelevant components. Attention-based visualization validates that predictions are guided by meaningful spectro-temporal patterns, supporting both interpretability and deployment readiness.

V. Conclusion

This study presented a compact and interpretable framework for speech-based gender recognition by systematically analyzing key factors that influence model reliability and robustness. Through focused experimentation, it was observed that appropriate acoustic representations and sufficient temporal context play a critical role in preserving gender-specific vocal characteristics. The integration of an attention mechanism

proved effective in selectively emphasizing informative spectro-temporal regions, leading to improved discrimination while maintaining a lightweight architecture. Robustness analysis under noise and channel mismatch conditions highlighted the model's suitability for real-world deployment scenarios, where recording environments and devices are often heterogeneous. Additionally, the incorporation of explain ability through attention visualization enhanced transparency, enabling a clearer understanding of the decision-making process. Overall, the proposed approach balances accuracy, robustness, and interpretability, offering a practical solution for speech-based gender recognition and establishing a strong foundation for future extensions involving cross-lingual data, fairness-aware modeling, and multi-attribute paralinguistic analysis.

References:

[1] M. M. R. Sindha and D. K. Rana, "Optimized artificial neural network for vocal gender recognition using a self-attention mechanism," ETRI Journal, 2024, doi: 10.4218/etrij.2024-0608.

[2] E. Yücesoy, "Gender recognition based on the stacking of different types of hybrid features created from speech," Applied Sciences, vol. 14, no. 15, Art. no. 6564, 2024, doi: 10.3390/app14156564.

[3] E. Yücesoy, "Automatic age and gender recognition using ensemble models on speech datasets," Applied Sciences, vol. 14, no. 16, Art. no. 6868, 2024, doi: 10.3390/app14166868.

[4] E. Yücesoy, "Speaker age and gender recognition using 1D and 2D convolutional neural networks," Neural Computing and Applications, 2024, doi: 10.1007/s00521-023-09153-0.

[5] S. Mavaddati, "Voice-based age, gender, and language recognition based on ResNet deep model and transfer learning in spectro-temporal domain," Neurocomputing, vol. 580, Art. no. 127429, 2024, doi: 10.1016/j.neucom.2024.127429.

[6] H. A. Younis et al., "Creating the Hu-Int dataset: A comprehensive Arabic speech dataset for gender detection and age estimation of Arab celebrities," Biomedical Signal Processing and Control, vol. 96, Art. no. 106511, 2024, doi: 10.1016/j.bspc.2024.106511.

[7] L. Yue et al., "Advanced differential evolution for gender-aware English speech emotion recognition with optimal feature selection," Scientific Reports, vol. 14, 2024, doi: 10.1038/s41598-024-68864-z.

[8] L. Yue et al., "Gender-driven English speech emotion recognition with genetic algorithm optimization and Fisher score," Biomimetics, vol. 9, no. 6, Art. no. 360, 2024, doi: 10.3390/biomimetics9060360.

[9] A. Garain et al., "GRaNN: Feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals," Neural Computing and Applications, vol. 34, no. 17, pp. 14463–14486, 2022, doi: 10.1007/s00521-022-07261-x.

[10] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," Multimedia Tools and Applications, vol. 81, no. 3, pp. 3535–3552, 2022, doi: 10.1007/s11042-021-11614-4.

[11] J. Radha and N. Gowrisankari, "Speech gender classification based on spectral and prosodic features with deep learning," International Journal of Speech Technology, 2023, doi: 10.1007/s10772-023-10039-8.

[12] J. Trawicki and M. Żyła, "Gender classification based on emotional speech: Deep learning and feature learning perspectives," International Journal of Speech Technology, 2024, doi: 10.1007/s10772-024-10090-z.

[13] A. Guerrieri et al., "Gender identification in a two-level hierarchical speech emotion recognition system," Sensors, vol. 22, no. 5, Art. no. 1714, 2022, doi: 10.3390/s22051714.

[14] L. M. Zhang et al., "A deep learning method using gender-specific features for speech emotion recognition," Sensors, vol. 23, no. 3, Art. no. 1355, 2023, doi: 10.3390/s23031355.

[15] D. Vlaj and A. Zgank, "Acoustic gender and age classification as an aid to privacy-preserving speech processing," Mathematics, vol. 11, no. 1, Art. no. 169, 2023, doi: 10.3390/math11010169.

[16] E. H. Alkhammash, "A hybrid ensemble stacking model for gender voice recognition," Electronics, vol. 11, no. 11, Art. no. 1750, 2022, doi: 10.3390/electronics11111750.

[17] D. Rizhinashvili et al., "Gender neutralisation for unbiased speech synthesising," Electronics, vol. 11, no. 10, Art. no. 1594, 2022, doi: 10.3390/electronics11101594.

[18] A. De Cario et al., "Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition," Engineering Applications of Artificial Intelligence, vol. 117, Art. no. 105651, 2023, doi: 10.1016/j.engappai.2022.105651.

[19] P. Taran, G. B. Kumar, and V. M. Suresh, "Speaker gender identification for voice assistants under device and channel variability," Applied Acoustics, vol. 205, Art. no. 109271, 2023, doi: 10.1016/j.apacoust.2023.109271.

[20] S. Shagi and A. S. M. Moin, "A comparative analysis for gender recognition using acoustic features and machine learning in real-world speech," Applied Acoustics, vol. 190, Art. no. 108392, 2022, doi: 10.1016/j.apacoust.2021.108392.

[21] A. A. Bhat, R. K. Garg, and S. Jain, "A deep learning approach for speech-based gender classification using spectral representations," Computer Systems Science and Engineering, 2024, doi: 10.32604/csse.2023.046730.

[22] C. Lisetti et al., "Identity, gender, age, and emotion recognition from speech using deep neural representations," Cognitive Computation, 2024, doi: 10.1007/s12559-023-10241-5.

[23] Y. J. Javid, R. Ahmed, and A. Ali, "Voice-based gender recognition using attention-based deep learning with multilingual speech signals," Wireless Communications and Mobile Computing, vol. 2022, Art. no. 4444388, 2022, doi: 10.1155/2022/4444388.

[24] G. De Simone, L. Greco, A. Saggese, and M. Vento, "Integrating visual and audio cues for emotion and gender recognition: A multi modal and multi task approach," Information Fusion, 2025, doi: 10.1016/j.inffus.2025.104071.

[25] M. Markitantov and O. Verkholyak, "Occlusion-robust audiovisual gender recognition and age estimation using attention mechanisms," Expert Systems with Applications, 2025, doi: 10.1016/j.eswa.2025.127473.

[26] O. H. Anidjar, R. Marbel, and R. Yozevitch, "An objective gender classification evaluation methodology for speech," Scientific Reports, 2025, doi: 10.1038/s41598-025-99011-x.

[27] Y. Hu, H. Zhang, and X. Li, "Gender-sensitive speech emotion recognition: A deep learning approach with robust feature fusion," Scientific Reports, 2025, doi: 10.1038/s41598-025-14016-w.

[28] C. Kirchhuebel, H. Jones, and A. Simpson, "Voice carries gender diversity: Classification limits of binary gender recognition from speech," Royal Society Open Science, 2025, doi: 10.1098/rsos.251193.

[29] S. Puri and V. Baghel, "Voice-based gender recognition with HeatMap analysis through machine learning," SN Computer Science, 2025, doi: 10.1007/s42979-025-03892-8.

[30] S. Yıldırım and İ. Bingöl, "Metaheuristic approaches to enhance voice-based gender classification and age estimation," Applied Sciences, vol. 15, no. 23, Art. no. 12815, 2025, doi: 10.3390/app152312815.