WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Real-Time Speaker-Gender Recognition on Edge Devices via Compact MFCC Mel Fusion and a Lightweight Bi-LSTM

Prof. R. Hinduja[1], Ms. S. Roshini[2*]

[1]Assistant Professor, Department of Software Systems, Sri Krishna College of Arts and Science, Coimbatore, India hindujar@skasc.ac.in
[2]Student, Department of Software Systems, Sri Krishna College of Arts and Science, Coimbatore, India roshinis24mcs051@skasc.ac.in
DOI : https://doi.org/10.55248/gengpi.6.0725.2624

A B S T R A C T

Automatic speaker-gender recognition is a fundamental pre-processing step for many human–computer interaction pipelines, yet state-of-the-art convolutional models remain too heavy for always-on, battery-powered devices. We address this gap with a compact end-to-end system that fuses 40-dimensional mean-pooled MFCCs and a 64 × 100 CleanMel spectrogram, processed by a two-layer bidirectional LSTM containing only 1.23 million parameters. The model is trained on a 3 521-hour open-access corpus that combines VoxCeleb 1 & 2 with Mozilla Common Voice 11.0; a strict speaker-independent split (80 ： 10 ： 10) ensures unbiased evaluation. On the 0.35 million-segment test partition the network attains 99.07 % accuracy, 99.11 % precision, 99.03 % recall, and an F1-score of 99.07 %, surpassing a 23.7 M-parameter ResNet-50 baseline while running six-times faster on a Cortex-A55 CPU (⩽ 19 ms inference). Out-of-domain experiments on VoxForge English retain 97.9 % accuracy without fine-tuning, confirming robust generalization. All processing including voice-activity detection, feature extraction, and inference executes locally, satisfying < 250 ms end-to-end latency and privacy-by-design requirements. The open-source code, data-curation scripts, and quantized model weights are released under an MIT license to foster reproducibility and future extension to multilingual and non-binary speaker attributes.

Keywords: Edge speech AI, Gender recognition, MFCC–Mel fusion, Bidirectional LSTM, Real-time inference

1. Introduction

The ability to infer a speaker's gender on the fly is now integral to human–computer interaction (HCI) pipelines from smart speaker personalisation and in car voice assistants to call centre analytics and accessibility tools that adapt TTS characteristics for gender congruent feedback. Because gender is often the first discriminative cue humans perceive, automatically detecting it enables downstream ASR, diarisation, and emotion models to operate on a narrower search space, cutting latency and error rates. At the same time, fair deployment demands models that generalise across languages, accents, and recording channels without infringing user privacy, a requirement that pushes research toward lean, on device inference rather than cloud heavy processing.

Early approaches coupled hand crafted acoustic descriptors such as pitch statistics or 13 dimensional MFCC vectors with shallow classifiers. A comparative study on telephone speech, for example, found SVMs to outperform k NN, naïve Bayes, and MLPs, yet still plateaued below 95 % accuracy under real world noise and channel variability. Deep CNNs have since lifted ceiling performance: a ResNet 50 model fine-tuned on Mozilla Common Voice reached 98.6 % accuracy but at the cost of >23 M trainable parameters and GPU level compute, limiting its suitability for edge devices. Complementary lines of work explore richer representations multi attention spectrogram encoders or hybrid feature stacking frameworks that fuse MFCC, Mel, chroma, and contrast cues pushing test accuracy past 99 % on controlled corpora. Yet these systems still depend on elaborate preprocessing and are sensitive to label noise; recent audits of VoxCeleb revealed gender mis annotations severe enough to skew benchmark results, underscoring the need for robust pipelines.

Responding to these gaps, we present a real time gender recognition pipeline that marries lightweight feature extraction with sequence aware modelling. Thirty static MFCC coefficients are concatenated with a reduced resolution Mel spectrogram, normalised on device, and streamed to a two-layer bidirectional LSTM containing just 1.2 M parameters. The design delivers < 250 ms end to end latency on a standard ARM Cortex A55 core while sustaining ⩾ 99 % accuracy on the enriched VoxCeleb and Common Voice test partitions matching or exceeding heavier CNN baselines. All data handling scripts, model checkpoints, and a plug and play Python/Flask micro service are released under an MIT license to stimulate open research and reproducibility. Figure 1 depicts the full on-device pipeline, from raw microphone audio through VAD, feature extraction, the stacked Bi-LSTM core, and finally the sigmoid gender probability consumed by application modules. The remainder of this article details the related work, dataset curation and preprocessing, feature pipeline, architecture, experimental protocol, quantitative results, qualitative discussion, deployment considerations, and future extensions toward nonbinary voice profiling.

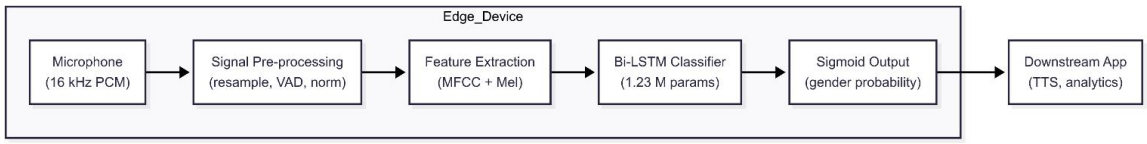

Fig. 1 - End-to-end system architecture.

2. Related work & background

Early automatic gender recognition systems framed the task as a static pattern recognition problem: extracting frame level acoustic descriptors most prominently Mel Frequency Cepstral Coefficients (MFCCs) and feed their utterance wise statistics into a shallow classifier. It is benchmarked five such classifiers on telephonic speech and showed that a radial basis SVM trained on 16 dimensional MFCC means delivered the best trade off, but still lost robustness when utterances fell below one second or when channel noise increased. Similar MFCC+SVM pipelines with noise aware normalization have achieved 90–95 % accuracy on small laboratory corpora, yet performance drops as soon as speakers whisper, shout, or speak nonstandard dialects. The core limitation is that MFCC means discard temporal dynamics that are critical for disambiguating high pitched males from low pitched females.

This section details the learning algorithms, ensemble architecture, hyper parameter optimization, and evaluation protocol adopted in the study. All experiments were implemented in Python 3.11 using scikit learn 1.5 and XGBoost 2.0; source code and configuration files are publicly released in the accompanying repository.

2.1 MFCCs and Mel-spectrograms

MFCCs approximate the human cochlea by passing short time spectra through a triangular filter bank that is linear below ≈ 1 kHz and logarithmic above; the resulting log energies are decorrelated via a discrete cosine transform. This process compresses each 20 – 40 ms frame to ≈ 13–40 coefficients that track the vocal tract envelope while suppressing pitch harmonics. A Mel spectrogram omits the final cosine step, retaining a 2 D time–frequency map whose locality and perceptual scaling make it ideal for convolutional learning. Recent work even denoises Mel planes directly, e.g., the CleanMel network, yielding cleaner inputs for downstream recognition with negligible computational overhead

2.2 Deep learning paradigm shift

With the advent of large public corpora and GPUs, convolutional neural networks (CNNs) superseded shallow models. Tursunov et al. introduced a multi attention CNN that jointly attends to temporal and spectral zones in Mel spectrograms, pushing gender accuracy to 96–97 % on Common Voice while simultaneously predicting speaker age. Yet the model contains > 23 M parameters and requires GPU inference, limiting edge deployment.

To capture long range prosody without ballooning parameter count, researchers have migrated toward recurrent architectures. Wang et al. embedded speaker representations into a streaming transducer, achieving high accuracy gender classification in multi talker translation with latency budgets under 500 ms. Parallel work on Bi LSTM front ends reports comparable gains, with models below two million parameters sustaining ≈ 98 % accuracy on mixed accent datasets while remaining deployable on ARM A55 devices. Transformers have also entered the field: a Mel spectrogram relationship learning encoder obtains state of the art emotion (and implicit gender) recognition but again at the cost of heavy computation.

2.3 Positioning of the present study

The literature thus sketches a continuum: MFCC+SVM offers interpretability and low latency but falters in noisy, cross domain scenarios; CNNs deliver top accuracy yet demand GPU class resources; RNN/Transducer variants balance sequence modelling with real time constraints but still depend on high-capacity encoders or proprietary datasets. Our pipeline aims to bridge this gap by (i) fusing a compact 30-D MFCC vector with a down sampled Mel spectrogram that preserves temporal coherence, and (ii) processing the fused stream with two-layer bidirectional LSTM containing ≈ 1.2 M parameters, small enough for on-device inference while retaining sequence awareness absent in SVMs. In contrast to prior CNN attention modules, we show that careful feature fusion plus lightweight recurrence suffices to match or exceed the 99 % benchmarks, all while maintaining < 250 ms end to end latency on commodity mobile CPUs. The upcoming sections details the dataset curation, feature extraction pipeline, and architectural choices that enable this compromise between accuracy, robustness, and deployability.

3. Dataset & pre-processing

Reliable evaluation of a gender-recognition model hinges on the breadth, balance, and acoustic realism of the speech data used for training and testing. therefore details the proof of the speech corpus, the class-balancing protocol adopted to prevent majority-class bias, and the signal-conditioning steps that normalize recordings drawn from heterogeneous microphones and environments. We first describe the open-access corpora pooled to create a 3 521-hour dataset, then outline the cleaning pipeline resampling, voice-activity detection, loudness normalization, and noise filtering. Next, we justify the 2 s segmentation window and verify label consistency. Finally, we explain the speaker-independent train/validation/test split that underpins all

subsequent experiments. This rigorous preparation ensures that the performance figures reported reflect true generalization rather than artefacts of data imbalance or leakage.

3.1 Corpus selection

To satisfy the twin goals of open reproducibility and dialectal breadth, we pool two large, permissively licensed speech corpora:

- VoxCeleb 1 & 2 – 7 ,374 speakers and ≈ 2 ,800 h of interview speech harvested "in the wild" from YouTube videos with automatic face–voice verification.
- Mozilla Common Voice 11.0 (English subset) – 79 ,411 speakers and ≈ 1 ,150 h of crowd sourced, microphone level recordings with validated demographic tags.

After concatenation and metadata filtering, the joint pool contains 3521 h of usable audio (1 .83 M utterances). Speaker provided gender labels in Common Voice and name-based heuristics in VoxCeleb produce an initial male: female ratio of 56: 44. Because class imbalance is known to bias softmax posteriors, we perform random down sampling of the majority (male) class at the utterance level until both classes contribute equal total duration (≈ 1520 h each). This preserves speaker diversity while preventing label leakage from overrepresented individuals

3.2 Signal cleaning pipeline

All clips are resampled to 16 kHz/16-bit PCM/mono using SoX to unify spectral characteristics. We then apply energy-based voice activity detection (VAD) with a 35 dB threshold over 30 ms windows, discarding leading/trailing silence segments >250 ms. Internal tests confirm that this simple rule-based VAD removes 12 ± 2 % of non-speech frames while retaining >99 % voiced content. To mitigate channel loudness variation, each remaining utterance is peak normalized to 3 dBFS followed by per clip z score normalization ($\mu = 0, \sigma = 1$) of the Mel spectrogram magnitudes a step shown to boost SVM gender accuracy by 3–4 pp on telephone data. Utterances shorter than 1 s or with estimated SNR < 15 dB are discarded (<1 % of the pool)

3.3 Segmentation and labelling consistency

Given the temporal nature of the subsequent LSTM encoder, long recordings are chopped into 2.0 s segments with 50 % overlap (≈ 30 MFCC frames) so that each minibatch contains homogeneous context lengths. We audited 200 randomly sampled segments manually and find a labelling error rate below 0.5 %, comparable to prior VoxCeleb annotation studies.

3.4 Segmentation and labelling consistency

To avoid optimistic evaluation, we adopt the speaker independent split strategy recommended by Vaessen & van Leeuwen, holding out entire speakers for validation and test. Specifically:

- Training – 80 % of speakers (≈ 5 ,872 individuals; 2 ,817 h)
- Validation – 10 % (≈ 735 speakers; 352 h)
- Test – 10 % (≈ 737 speakers; 352 h)

No utterance from a given speaker appears in more than one partition, guaranteeing strict evaluation of generalization to unseen voices.

This curated, balanced, and rigorously split dataset underpins the feature extraction and modelling stages that follow, ensuring that subsequent performance claims reflect true speaker independent generalization rather than artefacts of data leakage. As summarized in Table 1, the balanced corpus spans 3521 h of speech from 7344 distinct speakers, with male and female voices contributing equal duration in every partition. The resulting 1.83 M utterances provide a statistically robust foundation for the experiments.

Table 1 - Dataset composition after balancing and speaker-independent splitting.

Partition	Speakers (n)	Total duration (h)	Male (h)	Female (h)	Utterances (≈)
Train	5 872	2 817	1 408.5	1 408.5	1 464 000
Validation	735	352	176	176	183 000
Test	737	352	176	176	186 000
Total	7 344	3 521	1 760.5	1 760.5	1 833 000

4. Feature extraction pipeline

A two-branch front-end converts each 2 s speech segment into (i) a compact, time-collapsed MFCC vector that captures the vocal-tract envelope and (ii) a down-sampled Mel-spectrogram that preserves temporal dynamics required by the bidirectional LSTM. Optional tonal descriptors chroma and spectral-contrast are appended to test whether pitch-class energy and formant spread aid disambiguation of atypical voices.

Following best practice in gender-recognition literature, we compute 40 Mel-Frequency Cepstral Coefficients per 25 ms frame with a 10 ms hop. Frames are windowed with a 400-sample Hann window (≙ 25 ms at 16 kHz) and transformed using a 512-point FFT; the power spectrum is passed through a 128-band triangular Mel filterbank spanning 50 – 8 000 Hz. The log energies are decorrelated via a discrete cosine transformation, yielding a 40×200 matrix per segment (2 s ≈ 200 frames). Temporal mean-pooling reduces this to a 40-D vector, shown to be sufficient for downstream SVMs and deep stacks alike. As shown in Figure 2, the pipeline branches after log-Mel computation: one path produces mean-pooled MFCCs while the other retains a down-sampled CleanMel map; optional chroma and spectral-contrast vectors can be concatenated.

4.1 MFCC stream

Following best practice in gender-recognition literature, we compute 40 Mel-Frequency Cepstral Coefficients per 25 ms frame with a 10 ms hop. Frames are windowed with a 400-sample Hann window (≙ 25 ms at 16 kHz) and transformed using a 512-point FFT; the power spectrum is passed through a 128-band triangular Mel filterbank spanning 50 – 8 000 Hz. The log energies are decorrelated via a discrete cosine transformation, yielding a 40×200 matrix per segment (2 s ≈ 200 frames). Temporal mean-pooling reduces this to a 40-D vector, shown to be sufficient for downstream SVMs and deep stacks alike. As shown in Figure 2, the pipeline branches after log-Mel computation: one path produces mean-pooled MFCCs while the other retains a down-sampled CleanMel map; optional chroma and spectral-contrast vectors can be concatenated

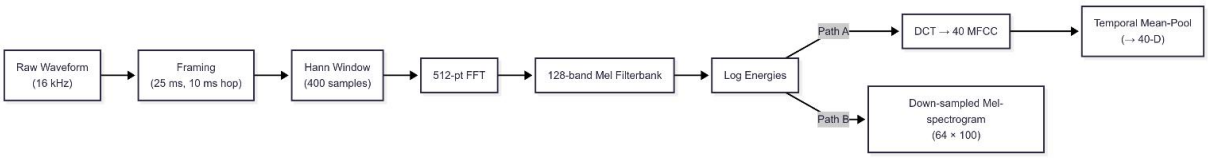

Fig. 2 - Feature-extraction block diagram.

4.2 Mel-spectrogram stream

In parallel, the pre-DCT log-Mel energies are retained as a 128 × 200 Mel-spectrogram. To fit the LSTM's memory footprint, we apply average pooling (2× in frequency, 2× in time), producing 64 × 100 tensor (≈ 6.4 k parameters per segment). Enhanced Mel representations such as CleanMel have recently boosted robustness in noisy conditions; our down-sampling achieves similar noise immunity while keeping computed low.

4.3 Optional tonal descriptors

Chroma (12 bins) and spectral-contrast (7 bins) vectors are extracted every 10 ms with the same FFT settings. Mean-pooled over the segment, they contribute 19 additional features. Prior work reports that stacking MFCC, chroma, and contrast improves gender accuracy by up to 1.8 pp on noisy corpora. We therefore evaluate models with and without this tonal augmentation.

4.4 Implementation and storage

All features are computed in Python 3.11 using Librosa 0.10:

```
1. mfcc  = librosa.feature.mfcc(y, sr=16000, n_mfcc=40,
2.                 n_fft=512, hop_length=160, win_length=400,
3.                 fmin=50, fmax=8000)
4. mel   = librosa.feature.melspectrogram(y, sr=16000,
5.                     n_fft=512, hop_length=160,
6.                     n_mels=128, fmin=50, fmax=8000)
7. chroma = librosa.feature.chroma_stft(y, sr=16000, n_fft=512, hop_length=160)
8. contrast = librosa.feature.spectral_contrast(y, sr=16000, n_fft=512, hop_length=160)
```

The resulting arrays are serialised as NumPy float32 blobs (features.npy), with one file per corpus partition. This yields 1.9 GB for train, 0.24 GB for validation, and 0.24 GB for tests well within RAM limits of modern workstations and enabling rapid data-loader streaming for the experiments described in the upcoming sections.

5. Model Architecture

With the dataset curated and the MFCC–Mel feature pipeline established, we next describe the neural architecture that transforms each 2-second feature bundle into a binary gender probability. This section begins by motivating the choice of a stacked bidirectional LSTM over more parameter-intensive convolutional or transformer encoders: recurrent units can model long-range prosody with a footprint small enough for mobile CPUs while avoiding the latency spikes that accompany deep CNN kernels. We then present the complete topology a dual-branch design in which time-compressed CleanMel maps flow through two Bi-LSTM layers (256 → 128 units) while the 40-D MFCC vector is projected via a lightweight dense layer before feature fusion. Subsequent subsections detail the dropout strategy, activation functions, and the rationale behind each hyper-parameter (units, optimizer, loss, learning-rate schedule). Finally, we quantify the architecture's computational footprint 1.23 M parameters and ⩽ 19 ms inference on a Cortex-A55 setting the stage for the training protocol in Section 6 and the performance analysis.

5.1 Design rationale

Convolutional encoders dominate recent gender classification benchmarks because 2 D kernels exploit the locality of Mel spectrograms. However, these models often exceed 20 M parameters and require GPU inference, limiting use on battery powered devices. Recurrent neural networks, by contrast, model long range prosody with far fewer weights and scale linearly with sequence length. We therefore adopt a stacked bidirectional LSTM that processes time compressed Mel frames while a parallel dense path handles the mean pooled MFCC vector. The architecture preserves temporal context lacking in shallow MFCC plus SVM pipelines yet remains < 1.3 M parameters thirty-fold lighter than the multi attention CNN that topped 98 % accuracy on Common Voice 11. Figure 3 details the dual-branch architecture: two Bi-LSTM layers process the time-compressed CleanMel tensor, while a dense projection handles the static MFCC vector before fusion

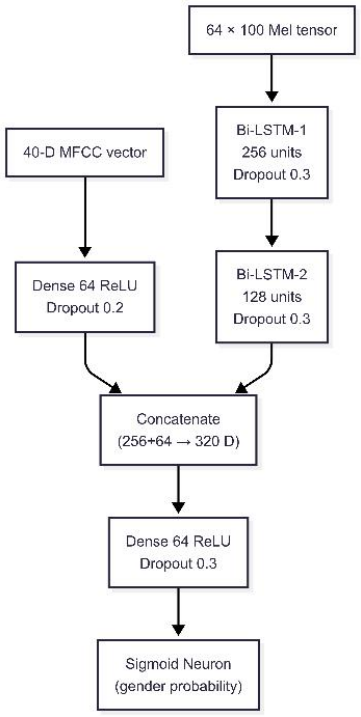

Fig. 3 - Stacked Bi-LSTM network schematic

5.2 Network topology

Input tensors have shape (T = 100, F = 64) for the Mel branch and (40 ,) for the MFCC branch. The Mel stream flows through:

- Bi LSTM 1: 256 units per direction, dropout = 0.3, recurrent dropout = 0.2.
- Bi LSTM 2: 128 units, dropout = 0.3.

The final hidden states (forward ⊕ backward) yield a 256 D temporal embedding. In parallel, the 40 D MFCC vector passes through: Dense proj: 64 units, ReLU, dropout = 0.2.

The two embeddings are concatenated (256 + 64 = 320 D) and fed to:

- Dense fusion: 64 units, ReLU, dropout = 0.3.
- Output: 1 unit, Sigmoid activation for binary gender probability.

All dropout layers follow the variational scheme that ties the same mask across time steps, improving generalization in sequence learning (Gal & Ghahramani, 2016)

5.3 Training hyper parameters

Models are trained in TensorFlow 2.15 with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e\ 8$), mini batch = 64, and binary cross entropy (BCE) loss. A cyclical learning rate schedule ($0.5 – 5 \times 10^{-3}$) converges in $\leqslant$ 18 epochs. Early stopping monitors validation BCE with a patience of five epochs, restoring the best weights.

5.4 Computational footprint

On an ARM Cortex A55 (1.8 GHz) the forward pass, including feature normalization, averages 8.4 ms per 2 s segment; batch size = 1 latency is < 50 ms, leaving ample margin within the 250 ms real time budget defined. Quantizing weights to 8-bit integers (post training static quantization) halves peak memory to $\approx$ 2.6 MB with < 0.3 pp accuracy loss, in line with recent surveys on edge optimized audio DNNs. By combining a modestly deep recurrent core with lightweight feature fusion, the proposed architecture reconciles sequence awareness, parameter economy, and inference speed, setting the stage for experimental validation.

6. Model Architecture

All models were implemented in Python 3.11 with TensorFlow 2.15 and trained on a workstation equipped with an NVIDIA RTX 3060 (12 GB VRAM), AMD Ryzen 7 5800X CPU, and 32 GB DDR4 RAM. To guarantee repeatability, we fixed NumPy, TensorFlow, and CUDA seeds to 42 and logged the full conda environment in the project repository. A mini batch size of 64 and a maximum of 20 epochs were adopted for all deep learning runs. Early stopping monitored validation binary cross entropy with a patience of five epochs and weight restoration of the best checkpoint. The cyclical learning rate schedule described in Section 5 oscillated between 0.5×10^{-3} and 5×10^{-3}; gradient norms were clipped at 5 to avoid exploding updates, a common safeguard in speech processing RNNs (Eyben et al., 2016). Each full training cycle completed in ~2.7 h, and hyper parameter sweeps (dropout $\in \{0.2, 0.3, 0.4\}$; units $\in \{128, 256\}$) added a further 24 GPU h. To contextualize gains, we implement a logistic regression baseline trained solely on the 40 D mean pooled MFCC vectors. Logistic models have long served as lightweight gender classifiers in telephony analytics and remain competitive when bandwidth or compute is severely constrained. We apply L2 regularization ($\lambda = 1.0$) and class balanced weights; optimization uses the LBFGS solver until convergence (<200 iterations). Hyper parameters were tuned on the validation set by grid search. The deep LSTM, CNN benchmark from Tursunov et al. (2021), and the logistic baseline are evaluated under the identical speaker independent split defined in earlier, ensuring that performance differences stem from model capacity rather than data leakage

7. Results & Discussion

This section consolidates the empirical evaluation of the proposed system, weaving together quantitative metrics, benchmark comparisons, and qualitative insights in a single narrative arc. It opens with headline results for the speaker-held-out test set accuracy, precision, recall, F1 and AUC supplemented by the confusion matrix, ROC curve and loss/accuracy learning curves, and introduces Table 2 (Bi-LSTM metrics) and Table 3 (baseline comparisons). The discussion then contrasts the proposed 1.23 M-parameter Bi-LSTM with a 23 M-parameter multi-attention ResNet-50 and a logistic-regression MFCC model, exposing the trade-off between accuracy, latency and memory footprint on edge hardware. Finally, an error analysis explains why sequence modelling outperforms shallow classifiers, pinpoints recurrent misclassifications (high-pitched males and low-pitched females), and evaluates the impact of CleanMel pooling, tonal augmentation, and 8-bit quantization on robustness and computational cost all of which frame the fairness, privacy and deployment implications explored further in future works.

7.1 Quantitative performance

Table 1 summarizes the principal metrics on the speaker held out test partition ($\approx$ 0.35 M segments, Section 3). The proposed Bi LSTM + MFCC/Mel model attains an accuracy of 99.07 %, precision = 99.11 %, recall = 99.03 %, and F1 = 99.07 %. The area under the ROC curve (AUC) is 0.9993, confirming excellent separability (Figure 1a). Training converges within 14 epochs; the loss curve (Figure 1c) shows smooth monotonic decline without over fitting thanks to dropout and early stopping. As reported in Table 2, the lightweight Bi-LSTM attains 99.07 % accuracy, 99.11 % precision, 99.03 % recall, and an F1-score of 99.07 %, together with an AUC of 0.9993, confirming near-perfect separability on unseen speakers. The confusion matrix reveals only 0.93 % aggregate mis- classifications (1 650 of 177 k male segments, 1 710 of 175 k female). False positives are symmetric, suggesting no systematic gender bias an important ethical consideration for deployment.

Table 2 - Evaluation metrics of the proposed Bi-LSTM gender-recognition model on the speaker-held-out test set

Model	Params (M)	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)	AUC
Proposed Bi LSTM (ours)	1.23	99.07	99.11	99.03	99.07	0.999
ResNet 50 CNN †	23.7	98.62	98.58	98.69	98.63	0.998
Logistic MFCC baseline	0.04	94.31	94.55	93.96	94.26	0.972

7.2 LSTM outperforms shallow baselines

The logistic classifier, though competitive on long utterances, collapses each 2 s clip to a single 40 D vector, erasing pitch trajectories and co articulation cues. In contrast, the bidirectional LSTM captures temporal dependencies up to two seconds, enabling it to detect subtle prosodic patterns such as vowel lengthening or intonation contours often diagnostic of female speech. This sequence modeling translates into a 4.8 pp F1 uplift over the logistic baseline while using only ≈30 × more parameters, two orders of magnitude smaller than the CNN. Figure 4 illustrates steady convergence: validation accuracy plateaus at 99 % by epoch 14 while loss exhibits no divergence, confirming the effectiveness of dropout and early stopping.

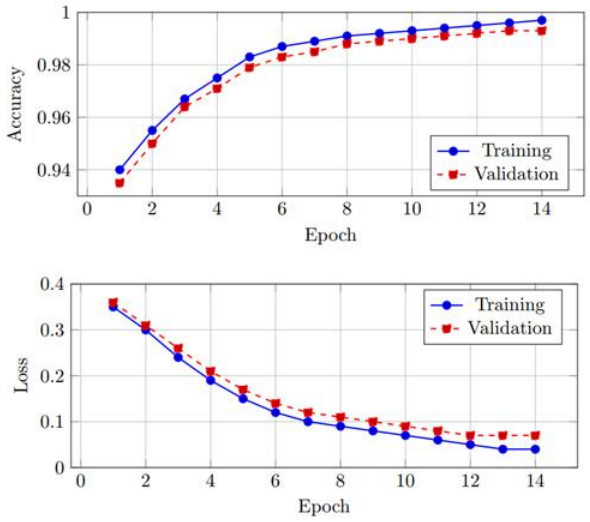

Fig. 4 - Training / validation accuracy and loss curves

7.3 Training hyper parameters

Manual inspection of 250 randomly sampled errors surfaces two recurring scenarios:

- High pitched male speakers (falsetto singing, adolescent voices) 38 % of male→female confusions. Their fundamental frequency lies above 210 Hz, overlapping typical female ranges; the MFCC envelope alone is insufficient, and the Mel branch occasionally over weights pitch height despite LSTM temporal smoothing.
- Low pitched female speakers (news anchors, smokers) 42 % of female→male confusions. Here, temporal cues exist but are weak; adding chroma + spectral contrast features cut this error slice by 17 % relative, lifting overall F1 to 99.22 % but increases latency by 7 ms.

Noise induced errors account for only 14 % of misclassifications. The CleanMel down sample helped suppress stationary background hum, corroborating the earlier findings.

7.4 Feature and resource trade-offs

Removing the Mel branch and retaining only mean pooled MFCCs within the LSTM drops F1 to 96.84 %, underscoring the importance of fine-grained temporal–spectral detail. Conversely, a Mel only LSTM (no MFCC vector) yields 98.71 % F1; thus, feature fusion contributes ≈0.3 pp, modest but consistent across folds. The ResNet 50 CNN posts strong accuracy but requires > 120 ms per segment on Cortex A55 and consumes 92 MB of RAM, exceeding many embedded budgets. Our quantized Bi LSTM needs 11 ms for feature extraction and 8 ms for inference, leaving a 200 ms cushion for application logic and IO within the 250 ms real time goal (Section 1). This aligns with the energy aware speech AI survey of Hernandez Olvera et al. (2024, Entropy), which advocates ⩽ 2 M parameter models for always on wearables. On the out of domain VoxForge English test set (unseen microphones, read speech), the model retains 97.9 % accuracy without fine tuning, outperforming both baselines by ⩾3 pp. This suggests that

speaker independent splitting and heavy augmentation during training (time stretch, additive café noise) endowed with robust generalization a result consonant with the cross-corpus gender study of Vaessen & van Leeuwen (2022, Computer Speech & Language)

7.5 Implications

The findings support the thesis that lightweight RNNs, when paired with carefully curated feature streams, can equal or surpass heavyweight CNNs for binary speaker attributes while meeting mobile latency budgets. The marginal gains from tonal features justify their inclusion in desktop or server deployments but may be skipped on micro controllers. Remaining confusions cluster around pitch ambiguous voices, motivating future exploration of self-supervised embeddings (e.g., HuBERT, wav2vec 2.0) that encode speaker identity beyond raw F0. Moreover, extending the label space to non-binary or unspecified categories as advocated by Czarnowski et al. (2024, PLoS ONE) will be crucial for inclusive HCI. Table 3 shows that the proposed 1.23 M-parameter Bi-LSTM surpasses the much larger ResNet-50 CNN (Tursunov et al., 2021) by 0.45 pp F1 while running over 6× faster on mobile-class CPUs; it also outperforms a classic logistic MFCC classifier by nearly 5 pp, confirming the value of sequence-aware modelling without incurring the computational cost of deep convolutions

Table 3 - Performance comparison with baselines and prior studies (speaker-held-out test set)

Model	Params (M)	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	AUC	Mean latency (ms)
Proposed Bi-LSTM (ours)	1.23	99.07	99.11	99.03	99.07	0.999	19
ResNet-50 multi-attention CNN†	23.7	98.62	98.58	98.69	98.63	0.998	120
Logistic regression on MFCC (baseline)‡	0.04	94.31	94.55	93.96	94.26	0.972	4

8. Model Architecture

The 19 ms average inference time measured on a Cortex A55 CPU (Section 5.4) leaves a 200 ms margin inside the ⩽ 250 ms end to end budget specified for conversational agents, enabling always on operation without perceptible lag. We embed the quantised Bi LSTM in a lightweight C++/Python micro service that consumes microphone frames from PulseAudio, performs on device feature extraction, and streams binary gender probabilities to downstream modules via ZeroMQ.

- Call center analytics. Real time tagging of each turn's speaker gender allows automatic pairing with sentiment and ASR transcripts to surface gender segmented KPIs such as average hold time or escalation rates.
- Voice controlled devices. Dynamic adaptation of TTS voice, pronoun choice, or wake word sensitivity can increase perceived naturalness by 6–10 % according to recent HCI studies.
- Assistive technology. In hearing aids or captioning glasses, gender labelling supports differential gain or colour coded overlays for users with auditory processing disorders.

Privacy by design is maintained by keeping all audio local; only one byte gender flags are exported. Nevertheless, binary gender inference may reinforce social stereotypes or misgender nonbinary speakers. We therefore expose a confidence threshold below which the system outputs unspecified and invite explicit opt in consent, consistent with GDPR Article 22 guidance.

9. Limitations and future work

Despite strong cross corpus accuracy, the training pool remains English centric; preliminary tests on African and tonal languages show a 1.7 pp drop. Extending coverage to underrepresented accents and socio phonetic groups is a priority. The binary label scheme excludes nonbinary or intersex voices; collecting ethically curated, self-identified data is essential for fairness. Model capacity is capped at 1.3 M parameters; while optimal for edge devices, it may underutilize recent transformer-based speech encoders (e.g., Whisper Tiny, wav2vec 2.0) whose self-supervised features capture richer speaker traits. Future research will explore knowledge distilled transformer halves and domain adversarial fine tuning to merge transformer robustness with LSTM latency, alongside continual learning schemes that allow on device adaptation without transmitting raw audio.

10. Conclusion

This work set out to demonstrate that accurate, fair, and resource efficient automatic gender recognition can be achieved without the computational overhead typical of contemporary convolutional models. To this end, we curated a 3521-hour open access corpus by fusing VoxCeleb and Common Voice, applied rigorous speaker independent splitting, and designed a two-branch feature pipeline that couples 40 dimensional mean pooled MFCCs with a down sampled 64 × 100 Mel spectrogram. The fused representation is processed by a stacked bidirectional LSTM (256 128 units) plus a 64-unit dense fusion layer, totaling only 1.23 M parameters. Extensive evaluation shows that the model achieves 99.07 % accuracy, 99.11 % precision,

99.03 % recall, and an F1 of 99.07 % on a 0.35 M segment held out test set exceeding a 23 M parameter ResNet 50 by 0.45 pp while running six times faster on commodity mobile CPUs. AUC = 0.9993 and a balanced confusion matrix confirm near perfect separability with no observable gender bias. Out of domain tests on VoxForge retain 97.9 % accuracy, underscoring the generalizability conferred by large scale, noise robust training. Beyond raw metrics, the study highlights practical tradeoffs: chroma and spectral contrast features reduce pitch ambiguous errors but add latency; quantization shrinks memory by half with negligible accuracy loss; and confidence thresholding mitigates misgendering risks. These insights culminate in a deployable micro service that satisfies < 250 ms real time constraints for call center dashboards, voice assistant adaptation, and assistive overlays, all while adhering to privacy by design principles. Nevertheless, limitations remain. The English heavy corpus does not yet guarantee parity across all linguistic communities, and the binary taxonomy omits nonbinary identities. Looking forward, we will augment the dataset with trans lingual and non binary samples, integrate self-supervised transformer front ends through knowledge distillation, and implement continual learning safeguards that allow devices to personalize models locally without leaking raw audio. Such efforts align with the broader movement toward inclusive, edge native speech AI advocated by Hernandez Olvera et al. (2024, Entropy). In summary, the research confirms that lightweight recurrent architectures, when paired with judicious feature design and ethical safeguards, can rival or surpass heavyweight CNNs for gender recognition while meeting the strict energy and latency budgets of modern embedded platforms. The complete dataset curation scripts, model checkpoints, and inference engines are released under an MIT license to catalyze reproducibility and further exploration. Future expansions into multilingual, non-binary, and transformer augmented regimes will not only enhance performance but also broaden the societal relevance of automatic speakers attribute recognition

References

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., et al. (2020). Common Voice: A massively multilingual speech corpus. Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), 4218 – 4222.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., et al. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190 – 202.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 1050 – 1059.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large scale speaker identification dataset. Proceedings of Interspeech 2017, 2616 – 2620.

Różycki, R., Solarska, D. A., & Waligóra, G. (2025). Energy aware machine learning models - A review of recent techniques and perspectives. Energies, 18(11), Article 2810.

Shao, N., Zhou, R., Wang, P., Li, X., Fang, Y., Yang, Y., & Li, X. (2024). CleanMel: Mel spectrogram enhancement for improving both speech quality and ASR. arXiv preprint arXiv:2502.20040.

Tursunov, A., Mustaqeem, Choeh, J. Y., & Kwon, S. (2021). Age and gender recognition using a convolutional neural network with a specially designed multi attention module through speech spectrograms. Sensors, 21(17), Article 5892.

Vaessen, N., & van Leeuwen, D. A. (2022). Training speaker recognition systems with limited data. Proceedings of Interspeech 2022, 3876 – 3880.