Advances in Artificial Intelligence and Machine Learning; Research 5 (4) 4747-4764 Received 25-09-2025; Accepted 17-12-2025; Published online 24-12-2025

Extraction of Gender Specific Hidden Information From Head Related Transfer Function Using Machine Learning

Tinny Sawhney tinnysawhney@jammuuniversity.ac.in
DSP Lab, Department of Electronics,
University of Jammu

Parveen Kumar Lehana pklehana@gmail.com
DSP Lab, Department of Electronics,
University of Jammu

Corresponding Author: Parveen Kumar Lehana

Abstract

Unique biological and behavioral characteristics are used for biometric identification because they contain reliable subject dependent information. Although conventional modalities such as facial geometry, vocal tract anatomy, fingerprints, iris patterns, and to some extent; gait also provide discriminative capabilities, their use is limited because of acquisition complexity, invasiveness, and sensitivity to environmental conditions. The growing need of non-contact, privacy-preserving biometric systems, research attention has shifted toward acoustic signals inherently being subject dependent. In this context, the Head Related Transfer Function (HRTF) has proven to be a reliable auditory biometric feature. It is direction and frequency dependent filtering of sound by the head, pinnae, torso, and shoulders. HRTF captures three cues: interaural time difference (ITD), interaural level difference (ILD), and pinna-induced spectral shaping. These spatially dependent cues vary in accordance with the morphological structure of the ear and its surrounding region. Our hypothesis is that HRTF not only encode subject specific information, gender-specific information may also be hidden within the acoustic signatures resulting from anatomical distinctions among the subjects. Cranial geometry, pinna curvatures, and torso volume are responsible for generating subject dependent acoustic signatures in the form of Head Related Impulse Response (HRIR). In the present research, a three-stage framework including spectrographic analysis, parametric cepstral visualization, and deep learning based acoustic signal classification has been used. Spectral and cepstral analyses showed gender dependent trends. Male subjects had stronger low frequency energy distribution and higher spectral variability. Female subjects showed enhanced sensitivity to high frequency content. The HRIR of female subjects had smoother cepstral gradients as compared to that of male subjects. For further verification, a hybrid Convolutional Neural Network with Bidirectional Long Short-Term Memory (CNN-BiLSTM) model was used. The model provided gender classification accuracy above 82%, a mean ROC-AUC of 0.98 and F1-scores above 0.95. As the public HRIR datasets have small number of subjects, Leave-One-Subject-Out (LOSO) cross-validation strategy was also used to ensure complete subject independence between training and testing data. The results shows

Citation: Tinny Sawhney and Parveen Kumar Lehana. Extraction of Gender Specific Hidden Information From Head Related Transfer Function Using Machine Learning. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(4):263. https://dx.doi.org/10.54364/AAIML.2025.54263

that HRIR can effectively be used for extracting the gender specific information hidden inside the given transfer functions derived from the corresponding HRTF.

Keywords: Head-Related Transfer Function (HRTF), Head Related Impulse Response (HRIR), CNN-BiLSTM, Deep learning, Acoustic signals, Interaural Time Difference (ITD), Interaural Level Difference (ILD).

1. INTRODUCTION

Biometric systems use various physiological and behavioral data to distinguish human identity. Traditional modes including handwriting, hand geometry, gait, facial features, speech, iris, retina, and fingerprint each have both strengths and limitations [1–5]. Handwriting variability can be caused by emotional state [1], the limited discriminative power of hand geometry [2], and the limits of behavioral biometrics in examining gait-related environmental effects [3], exemplify the shortcomings of behavioral biometrics. Despite being extremely accurate, face recognition and fingerprint analysis rely on structured image acquisition and are susceptible to data quality problems [4, 6, 7]. Speech-based systems offer non-invasive analysis but depend on acoustic coherence with high sensitivity to channel orientation and filter effects associated with head-related filtering effects [8]. Recent research in deep learning has dramatically improved the extraction of biometric cues from speech and images, enabling strong modeling of gender-related features associated with sex-related anatomical differences of vocal folds, resonance paths, and articulatory structures [9–13]. Collectively, these data indicate that most biometric features intrinsically encode gender-specific morphology. Analogous to this in the auditory domain, the Head-Related Transfer Function (HRTF) has originated as a biometric marker with valid physiological bases. HRTF characterizes directional acoustic changes induced by the head, pinna, torso, and shoulders before sound gets to the eardrum [5]. These transformations produce three information features: interaural time difference (ITD), interaural level difference (ILD), and pinna-induced spectral notches, which are much more sensitive to listener morphology.

Classical and modern studies show that HRTFs are both individual-specific and gender-influenced. Measurements from simulators, psychoacoustic studies, and anatomical modeling confirm that variations in head circumference, pinna geometry, shoulder breadth, and torso length between males and females shape spectral filtering patterns and spatial cue distributions [14–22]. Databases and measurement frameworks, including near-field HRTF datasets and laser-based acquisition, further confirm morphological dependence in HRTF formation [23]. The anthropometric models [21, 24, 25], have also been used for gender determination, that rely on measurements of skeletal or internal anatomical structures. Although skeletal differences do influence HRTFs, directly measuring these anatomical parameters is not feasible in most practical scenarios.

Although traditional techniques such as MFCCs based Support Vector Machine (SVM) and logistic regression are prevalent for speech perception and loudness, but found to ignore high frequency and phase feature information that is important for spatial hearing and anthropometric cues, thus limiting their applicability for an extensive HRIR-based gender study. On the other hand, deep learning methodologies, specifically CNNs, DNNs, and hyper-conditioned models have been successfully applied for investigations related to HRTFs where anthropometric features and ear images are reported by researchers incorporating subject dependent features [14, 15, 21, 24, 25].

The scope of the present research is to explore if HRIRs inherently encode gender-specific information using a hybrid CNN-BiLSTM architecture. The model is expected to learn subtle temporal and spectral cues for gender identification directly from HRIRs. HRTFs are structurally defined by human morphology, the present research investigates if low-level acoustic features related to gender specificity are inherently present in these responses. The proposed method not only offers a contactless, privacy-preserving auditory biometric modality but also lays the foundation for future applications in spatial audio personalization, hearing-aid optimization, virtual and augmented acoustic environments, and computational auditory biometrics. The detail of HRTF is presented in the following section. followed by the. methodology and the results.

2. HEAD RELATED TRANSFER FUNCTION

Hearing perception in humans can only get accurate spatial auditory processing using the binaural pressure of the auditory sounds in reality, like localization of the sound sources, and localization of the reflections of the environment. The Head Related Transfer Function (HRTF) is a basic auditory descriptor for the humans. It is the mechanism by which sound interacts with the anatomical structures embedded within the precursors of the ear canal. It establishes the spectral and temporal changes in an acoustic wave as it carries it on its path, through its origin to the listener's ear, determined by the morphology of the head, pinnae, torso, and shoulders [26, 27]. The procedure for the measurement of HRTF is shown in **FIGURE 1** [14].

Figure 1: The schematic diagram showing the procedure for the measurement of HRTF [14].

Each individual demonstrates distinct HRTF profile shaped by cranial profile, auricular geometry, and body frame, providing a distinct "acoustic fingerprint" that underlie three-dimensional auditory

perception and recognition for identity. In the field of pure sound, sound waves propagate directly to the listener through both direct and reflected channels and are altered through scattering, diffraction and reflection by the body and head, thus causing pressure differences at the two ears, which are interpreted by the auditory system into spatial perception [28].

Psychoacoustic studies have indicated that there are several cues which are beneficial to directional localization: The interaural time difference (ITD), especially effective in frequencies lower than about 1.5 kHz [29]; the interaural level difference (ILD), important in frequencies higher than 1.5 kHz because of head shadowing [30]; and spectral cues related to frequency-dependent scattering and reflections, especially emitted from the pinnae above 5–6 kHz, leading to front–back and vertical localization [31]. Dynamic cues caused by the head motion improve localization accuracy and eliminate front–back ambiguities [32]. Spherical coordinates (r, θ, ϕ) that capture the position of sound sources are the r in a spatial sense, θ is the azimuth at the ground level and ϕ is an elevation at the ground scale. With stationary both the listener and sound source, the acoustic transmission from source to hearing is a linear time invariant (LTI) system where HRTF is defined as:

$$H_L(r, \theta, \phi, f, a) = \frac{P_L(r, \theta, \phi, f, a)}{P_0(r, f)} \tag{1}$$

$$H_R(r, \theta, \phi, f, a) = \frac{P_R(r, \theta, \phi, f, a)}{P_0(r, f)} \tag{2}$$

where P_L and P_R are the left and right ear sound pressures, and P_0 is the head center pressure on a free field. HRTFs vary with frequency (f), spatial position (r, θ, ϕ), and the listener (far-field HRTFs ($r > 1.0 - 1.2$ m)) having an independent distance and near-field HRTFs ($r < 1.0$ m) having distance dependent HRTFs. Additionally, HRTF can be decomposed [33], into a minimum-phase component, an all-pass component, and a linear-phase term as:

$$H(\theta, \phi, f) = H_{\min}(\theta, \phi, f).\exp(j\psi_{all}(\theta, \phi, f).\exp[-j2fT(\theta, \phi)] \tag{3}$$

When minimum-phase contribution from all-pass is negligible, HRTF can be simplified as a function of the product of its minimum-phase function and a linear phase (pure delay), a simplification referred as minimum-phase approximation and is valid for most instances below 10–12 kHz [34]. The all-pass term can introduce small error in the estimation of ITD. In [35], it is reported that at levels below 1.5 kHz, all aural phase contribution to interaural group delay differences is nearly frequency-independent and the approximation would not lead to error more than 30 μs, which is perceptually insignificant [36]. Physiologically, the HRTF contains the auditory cues of three elements ITD, ILD, and spectral filtering affecting the overall binaural hearing and spatial information conveyed by the sound. In addition to spatial localization, HRTFs encode rich spectral information that enables the auditory system to discriminate between sources and recognize familiar voices based on identity-related acoustic patterns.

Literature survey suggests that the HRTFs' contain enough spatial distinctive features for confirming biometric identity of the subjects. The spatial variation of HRTF profiles modifies the auditory characteristics of vocal timbre and speaker identity. With spatial localization of data and vocal input, one can interpret a scene. This in turn provides auditory scene decomposition by isolating concurrent voices, making its detection power in the multitalker environment much quicker and more accurate.

The description and modeling of HRTFs is the backbone of auditory science and spatial signal processing. Two fundamental approaches are used, including free-field recordings of sound prop-

agation under natural settings for capturing acoustic propagation and closed-field measurements in anechoic environments to remove reflections [37]. Experimental setups are designed which replicate human anatomy and are equipped with microphones to measure binaural responsivity. The acquired data is applied to the frequency domain to obtain detailed directional maps of spectral filtering behaviors. Personalizing HRTFs is crucial since generalized databases cannot accurately identify the individual. Individualized HRTFs (indirect measurement/taken to heart or through synthesis) include both finite element modeling (FEM) and boundary element methods (BEM) which provide better realistic auditory localization and identity cues [38].

Brinkmann et al. [39], reported that the dynamic binaural synthesis may provide realistic sounds for virtual auditory applications like generation of HRTF based acoustic signals for embedding spatial information. Such mismatched and generic HRTFs lead to spatial disorientation and reduced intelligibility due to inconsistency between genuine anatomical filtering and simulated acoustical cue. Additionally, HRTF-based perception is mediated by characteristics including reverberation, distance, and room configuration as they modify spectral characteristics critical for identity perception [40].

In HRTF modeling, some combination of machine learning and adaptive filtering techniques are emerging to be applied and to obtain identity-relevant features that can withstand noise and reverberation. Reliable HRTF representation should lead to the enhanced auditory realism in immersive and assistive technologies and facilitate the development of gender- and identity-specific digital modeling that is suitable the biometric applications of such technologies. It has been reported that subject differences in HRTF, which arise from differences in head size, pinna style and torso size, exhibit significant differences in spectral cues, which can be identified using deep learning networks [41, 42].

HRTF is a multidimensional signal between acoustics, anatomy, and cognition, and records both the physical transformation of sound and its perception. Its use is not limited to auditory localization, but can be further explored for gender recognition having application in neuroscience, spatial audio engineering, and emerging biometric technologies. Using advanced computational models like attention-based deep learning architectures, our work is related to explore latent, gender-dependent patterns within datasets, leading to intelligent, perceptually grounded systems that combine anatomical realism with cognitive precision.

3. MATERIAL AND METHOD

The proposed research analyses gender-specific information encoded into HRTFs. The methodology of the system employed for the proposed investigations is shown in **FIGURE 2**. HRTF in time domain is referred as Head Related Impulse response HRIR. For the investigations, RIEC HRTF dataset [43], developed at Tohoku University has been used. The measurements were taken in anechoic environment using probe microphones and spherical loudspeaker grid (**FIGURE 1**). The HRTF dataset used in the study consisted of the HRIRs from 105 subjects sampled at 48 kHz sampling rate in SOFA format. It comprises of 27 anthropometric males, 10 anthropometric females, 2 anthropometric subjects with unspecified gender, 15 actual males, 1 actual female, and 50 subjects with unspecified gender. For every subject, HRIRs are available at 865 spatial directions per ear, giving 1,81,650 HRIRs in the full dataset.

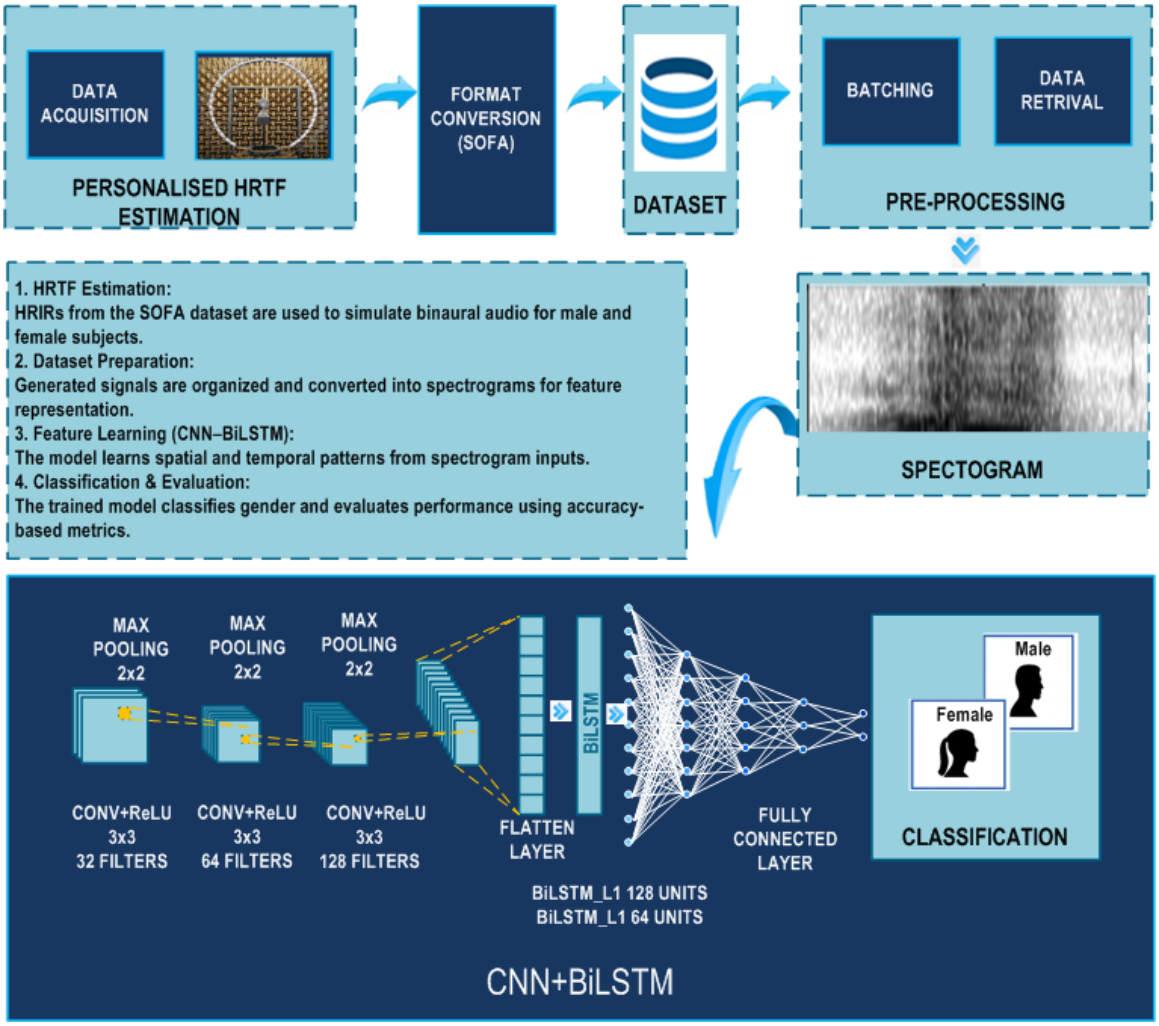

Figure 2: Block diagram of the proposed HRTF-based gender classification system, illustrating HRTF acquisition, SOFA-format conversion, spectrogram generation, and the CNN-BiLSTM network used for spatial–temporal feature learning and final gender classification.

Publicly available HRIR datasets have small number of subjects and inherent gender imbalance, which may hamper the gender identification accuracy. To address this issue, two different strategies were employed for the investigations. The first strategy, a set of 53 subjects (42 males and 11 females) was used giving 4,58,450 HRIRs ($53 \times 865 \times 2 \times 5$) after data augmentation using five noise levels with standard deviations from 0 to 0.005. The total available dataset is split into training (70%), testing (20%), and validation (10%) set with stratified shuffle split. To avoid class imbalance, the second strategy used Leave-One-Subject-Out (LOSO) technique with 22 subjects (11 males and 11 females) to rigorously enforce subject-independent evaluation by ensuring that all samples from a given subject were excluded from training and used only in testing.

HRIRs are first converted to a spectrogram representation, allowing CNN blocks to learn localized spectral-spatial features encoded in the signals, followed by a BiLSTM layer to capture contextual dependencies across frequency bins and to also capture long-range effects ignored by shallow classifiers. The investigations carried out are categorized as spectrographic analysis, parametric visualization, and deep learning based validation. The spectrographic and parametric analysis focuses on the interpretation of physical and perceptual differences of acoustic cues for the male and female subjects, while deep learning using CNN-BiLSTM model has been used for validation. The approach emphasises gender specific analysis providing quantitative validation of the auditory behavior of the subjects under consideration. The HRIRs were preprocessed in accordance with spatial metadata (azimuth, elevation and impulse response length). The Short Time Fourier Transform (STFT) was applied to examine each waveform produced by 1024-point FFT, Hanning window, 20 sample hop length. The obtained magnitude spectrograms were converted to logarithmic decibel scale and normalized to a constant [0,1] value. The spectrograms serve as a basis to investigate frequency patterns related to HRTF-induced filtering effects at different spatial orientations. For parametric analysis, spectral-envelope deviations along the spatial orientation were measured through cepstral distance based metrics. HRIRs from different azimuth and elevation angles were compared using a logarithmic cepstral distance metric. Three-dimensional cepstral surface plots were created using cepstral distances computed pairwise at each orientation for each ear. These parametric visualizations were deployed to compare patterns of spatial spectral variation in HRIRs for a systematic comparison of directional acoustic behavior.

The deep learning based classification is intended to detect gender-specific information inside HRTFs. The workflow is divided into four main phases i.e. HRTF estimation, feature extraction and preprocessing, dataset preparation, CNN-BiLSTM model training, testing, and evaluation of performance metrices phase. The binaural audio files were stored into hierarchical directories by gender and azimuth elevation configuration. Each file was annotated by gender and defined spatial metadata to ease stratified sampling while training, testing, and validating. This dataset was used as the basis for deep learning framework. The magnitude spectrograms (513×17) of the HRIR on decibel-scale were obtained and adjusted to a specific pixel scale using anti-aliasing interpolation. Here, 513 frequency bins correspond to a 1024-point FFT, retaining only the positive frequency components. HRIRs are inherently short, transient impulse responses, and results in 17-time frames only. The spectrograms were normalized to [0,1] to ensure uniform dynamic scaling. Mel-Frequency cepstral coefficients (MFCCs) were calculated for STFT based features to capture the perceptually relevant spectral envelope details with respect to their distance. Labels were homogenized for one-hot encoding and the resulted dataset was used for the further investigations.

The proposed architecture was used to capture spatial spectral and temporal features from HRTF-based spectrograms. The model comprises 3×3 kernel convolutional blocks, ReLU activation function, He-normal initialization, max-pooling, and dropout regularization rate of 0.25. The convolution module captures multi resolution spectral patterns while using the reduced dimensionality. The output feature maps were first flattened and then recombined in sequential form, which were subsequently processed by two stacked BiLSTM layers (128 units and 64 units), allowing modeling of temporal dependencies across spectral frames. The fully connected layers include dense layer (64 neurons) and softmax output layer (the two units are representative of each male and female class). The proposed model was trained using the Adam optimizer with a learning rate of 0.001, for 100 epochs and batch size of 32. Two callbacks were used during training to store the best model weights as an outcome of validation accuracy, and to reduce the learning rate when validation loss

reached plateau. During testing and validation, each audio sample was preprocessed and applied to the trained CNN-BiLSTM classifier to generate gender probability estimates. Outcomes of the models, confidence scores, and related metadata were stored in the CSV files for later analysis.

4. RESULTS AND DISCUSSION

The results of the investigations of spectrographic analysis, parametric visualization, and deep learning based classification for first strategy are detailed as follows.

The binaural signals, which are generated from the HRIRs, are recorded in a time-frequency format for spectrographic analysis. For example, **FIGURE 3** and **FIGURE 4**, show a comparative spectrographic representation of a male subject and a female subject, respectively. The figures show HRIRs waveforms, spectrograms at pair of different angles of elevation azimuth (E0A0, E0A90, E0A180, E0A270), and three-dimensional cepstral distance plots to investigate spatial sensitivity. HRIRs of the male subject (**FIGURE 3**) demonstrate clearly defined early reflection peaks and relatively elongated impulse response especially for the left-ear, indicating interaural path-length differences and more powerful head diffraction. The spectrograms show dominant low-to-mid frequency energy (~0.5–3.5 kHz), slower time decay, and wider spectral envelopes. This behavior correlates with a larger cranial size and more pronounced scattering effect, which improves spatial robustness at low frequency with smooth high-frequency localization cues. On the other hand, the female HRIRs (**FIGURE 4**) are characterised by sharper transient peaks and shorter decay periods, indicating small propagation delays and reduced low frequency attenuation. Thus, spectral resolution is improved and diffraction effects are reduced as more focused high-frequency content (greater than 5 kHz) is seen in the spectrograms.

The results of parametric visualization are shown in **FIGURE 5**. It shows a three-dimensional cepstral distance surfaces of six subjects where the first two columns correspond to male subjects (M001, M002, M003) and the last two columns correspond to female subjects (F004, F012, F018), shown separately for the left and right ears. It is clear from the plots that cepstral differences are different for male and female subjects with steeper surface gradients around lateral azimuths (90° and 270°). The differences indicate stronger spectral coloration, greater interaural asymmetry, and more complicated direction-dependent filtering. The differences may be attributed to head radius, widening of interaural spacing, and pinna-torso geometry. The peaks indicate stronger low-frequency diffraction and complex multi-path scattering effects for male subjects. The corresponding cepstral plots of female subjects have smoother topographies, less transitions, and consistent spectral behavior through angles of azimuth and elevation. The cepstral discontinuities are smaller, resulting in smaller peaks indicating relatively more cohesive high-frequency content as compared to the low cepstral region. Local high-frequency peaks at higher elevations are more pronounced with sharper pinna resonances and narrower acoustic pathways. Overall, **FIGURE 5**, shows that the male subjects are characterized by wider distances, spectral fluctuations, and spatial complexity of their envelopes and the female subjects exhibit compact, stable, high-frequency dominant, and high-frequency centered behavior. The observations align with the anatomical differences between male and female subjects.

In order to quantify the gender-based acoustic distinctions revealed in spectrographic and cepstral analyses, a deep learning-based classification framework was chosen as the third analytical phase of

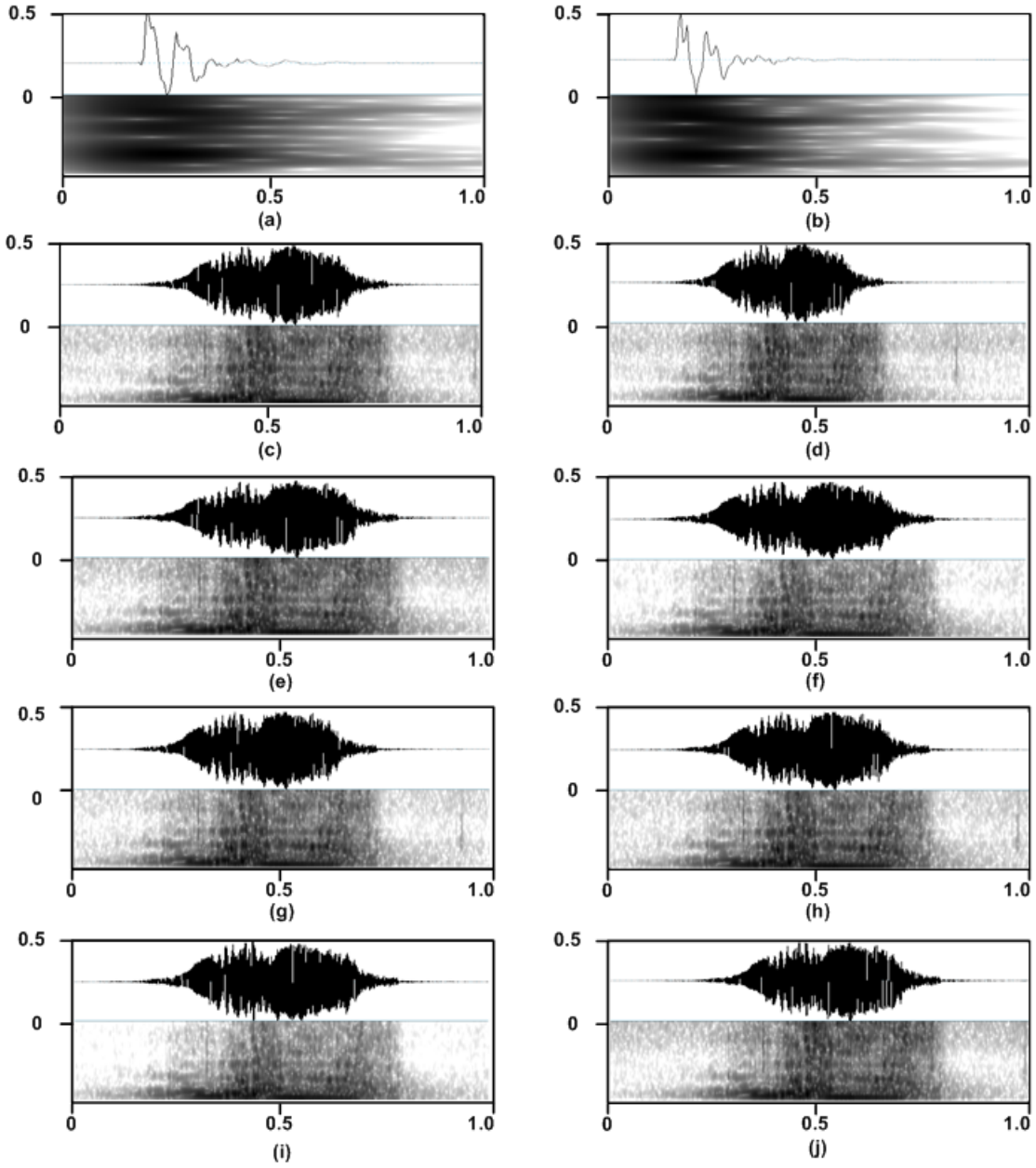

Figure 3: Comparative analysis of Head Related Impulse Responses (HRIRs), spectrograms, and cepstral distance maps for a male subject across multiple azimuth–elevation configurations. Subplots (a) and (b) show the time-domain HRIRs for the left and right ears, respectively. Subplots (c)–(j) illustrate the corresponding waveforms (upper panels) and time–frequency spectrograms (lower panels) for azimuth–elevation combinations of E0A0, E0A90, E0A180 and E0A270 for both ears.

the proposed system. The framework utilises the information extracted from the earlier phases and

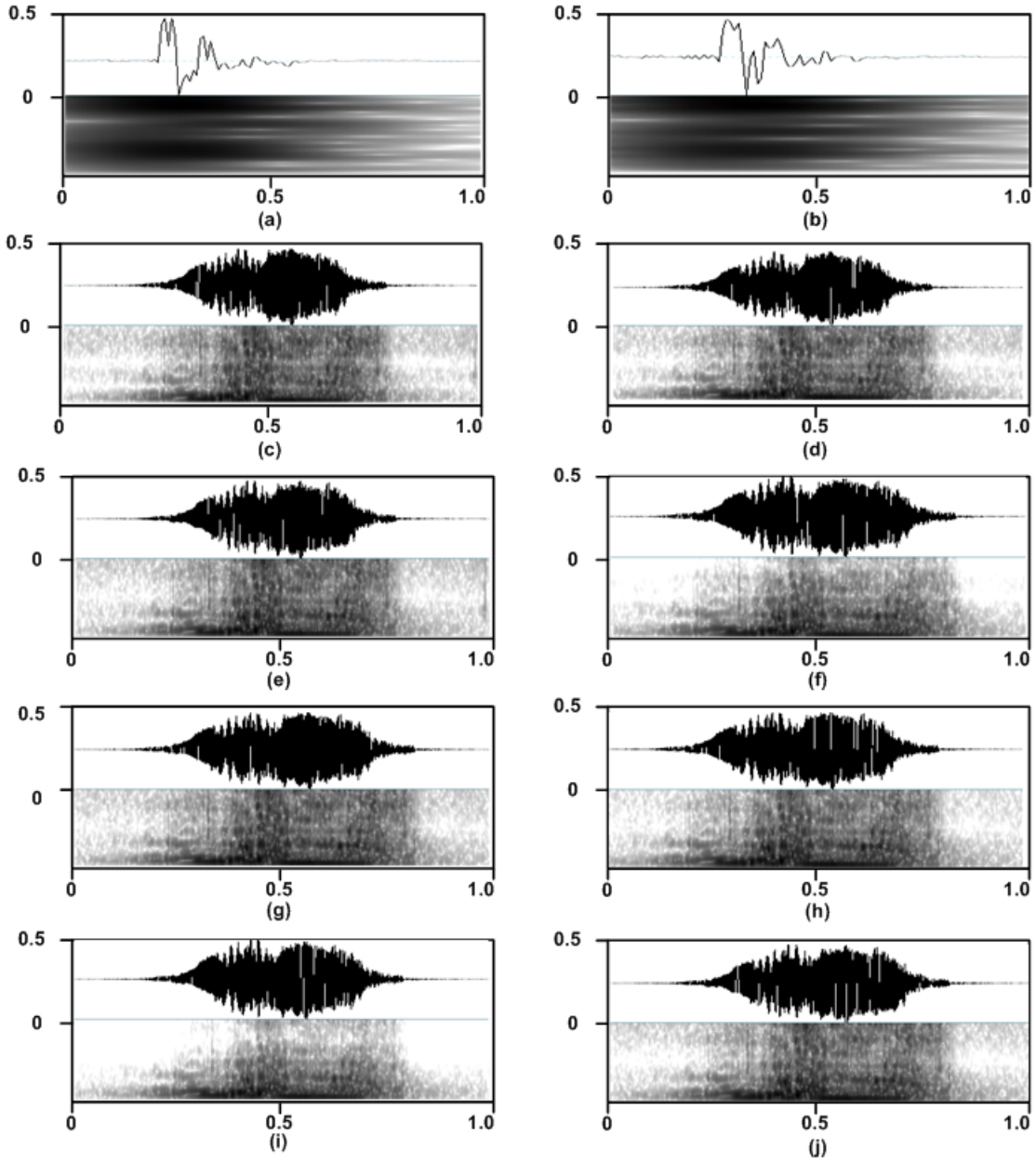

Figure 4: Comparative analysis of Head Related Impulse Responses (HRIRs), spectrograms, and cepstral distance maps for a female subject across multiple azimuth–elevation configurations. Subplots (a) and (b) show the time-domain HRIRs for the left and right ears, respectively. Subplots (c)–(j) illustrate the corresponding waveforms (upper panels) and time–frequency spectrograms (lower panels) for azimuth–elevation combinations of E0A0, E0A90, E0A180 and E0A270 for both ears.

discriminative capability of deep learning networks. The HRIRs were converted to spectrograms, providing spectral and temporal information embedded in the HRTFs. The spectrograms were used

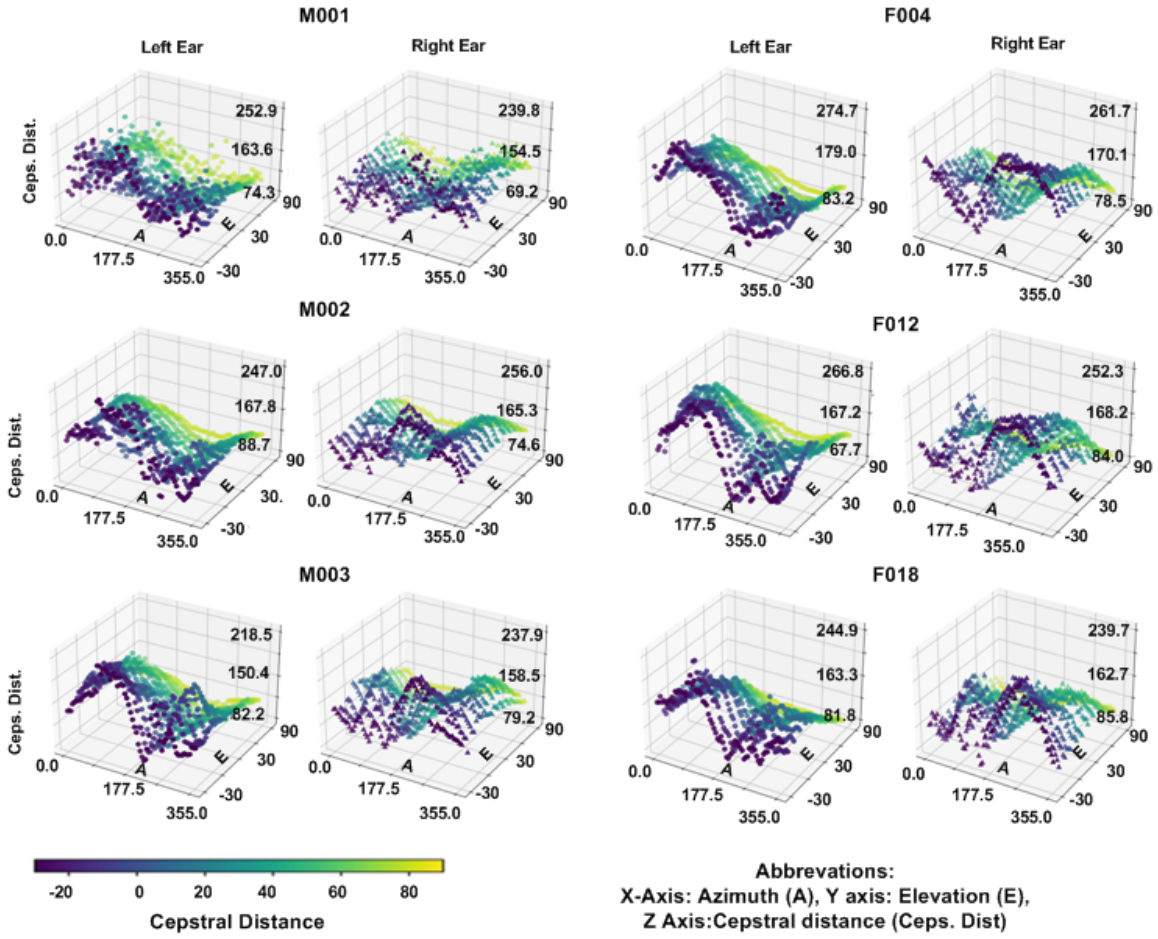

Figure 5: First two columns are the Cepstral Distance for Male Subjects M001, M002, M003 for both left and right ear whereas the third and fourth columns are Cepstral Distance for Female Subjects F004, F012, F018 for both left and right ear.

as input to a CNN-BiLSTM model. CNN layers capture spatially varying spectral content indicating the effect of pinna resonances, spectral notches, and diffraction induced modulations, while the BiLSTM layers utilise temporal dependencies and directional consistency along elevation azimuth dimensions.

The model was trained on labeled dataset that included both male and female HRTF responses in an effort to determine whether the observed spectral differences were sufficiently distinct to make them compatible with automatic classification. The elevation angle was fixed at 0° and azimuth angle varied from 0° to 315°. The training curves (**FIGURE 6)** show that the training accuracy improves gradually to 99%, while validation accuracy remains 94%, indicating strong generalization. Training loss dropped from 0.68 to 0.05 and validation loss remained stable at 0.25, showing effective optimization of the model and controlled overfitting achieved because of appropriate dropout and adaptive learning rate scheduling.

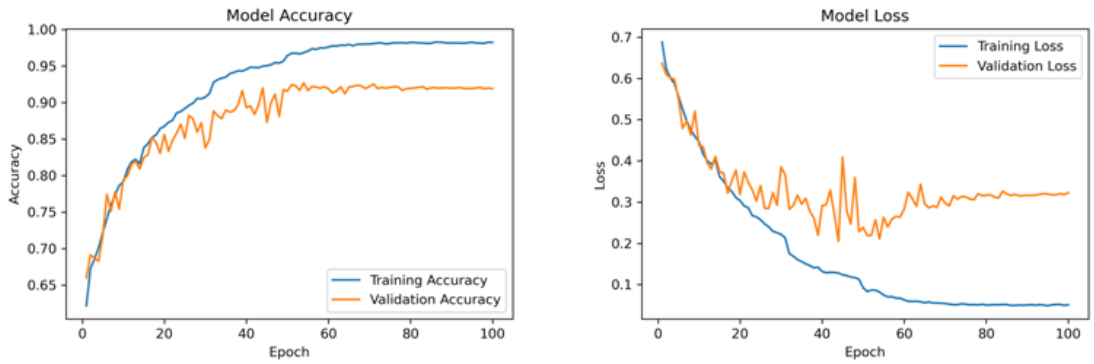

Figure 6: Training Accuracy and Loss curves.

The investigations showed that the CNN layers learned spatial spectral cues. The BiLSTM layers captured temporal information from spectrograms effectively. The testing results are shown in **FIGURE 7**. In frontal direction indicated by E0A0 (means elevation angle 0 and azimuth angle 0 in degrees), the resulting accuracy was 97%, precision as 0.97, recall as 0.97, and F1 as 0.97. ROC–AUC was obtained as 0.998, signifying great separability. Lateral orientations (E0A90 and E0A270) showed moderate performance (82.5–87.5% accuracy) owing to the more pronounced coloration, but still maintained considerable class distinctions. Rear orientation (E0A180) achieved 92.5% accuracy with strong MCC and AUC scores, indicating successful modeling of reflective and reverberant cues. The model achieved 100% accuracy for the oblique orientation (E0A315). Among all spatial configurations, the classifier showed higher accuracy above 82.5%, mean ROC–AUC around 0.984, and mean MCC as 0.868 indicating the existence of highly separable gender-specific acoustic features in HRTFs. The results are also *-*presented in **TABLE 1**. The model was able to capture invariant morphological cues and angle-specific spectral information in the HRTFs, showing robust generalisation. The model took about ∼12 s training time per epoch and < 0.05 s inference time per sample on a GPU-enabled system, substantiating its suitability for live perceptual modeling and sound classification tasks. Overall, the deep learning based validation shows that HRTFs represent directionally coherent, gender-specific signatures that can be effectively extracted by the proposed model.

In **FIGURE 7**, at the spatial position E0A315, the initial performance evaluation of gender classification using multiple HRIRs yielded an apparent accuracy of 100%. Such a result may be intuitive (at least initially), but it must be noted that HRIR samples for the same subject naturally exhibit spatial correlation. As a result, random splitting of HRIR data might result in subject-level information leakage, possibly overestimating model performance. These effects are even more critical in tasks of learning models based on HRIR, where acoustic signatures could be learned for specific subjects instead of generalizable gender-related cues. To overcome this limitation, the second strategy LOSO cross-validation was implemented and the experimental protocol was modified to ensure complete subject independence between training and testing data.. In each batch of LOSO, the HRIRs from 10 male and 10 female subjects ($2 \times 17{,}300$ HRIRs) were utilized for training purposes, and HRIRs from one male participant and one female subject resulting in 3,460 HRIRs were used for testing. By performing this sequence on all subject pairs, 11 independent LOSO evaluations were achieved.

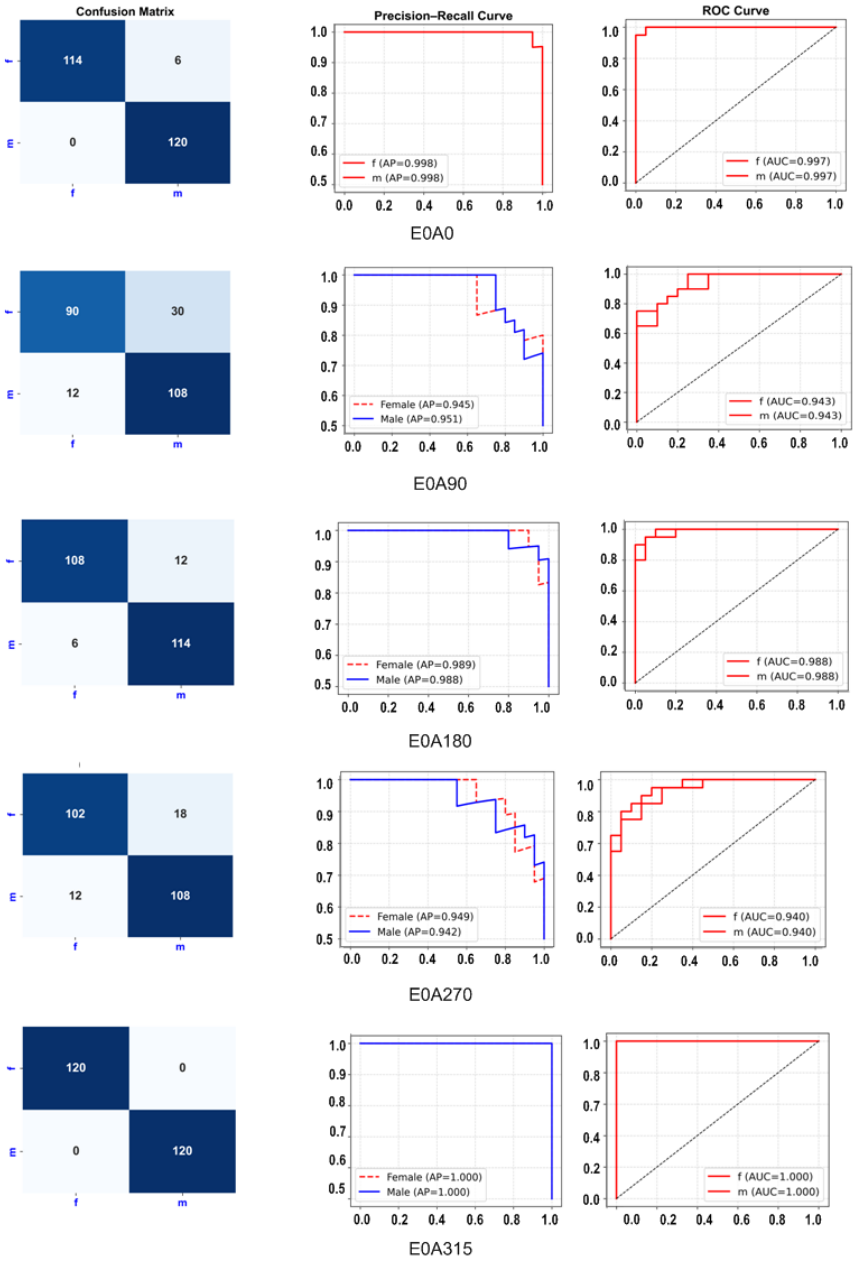

Figure 7: Performance evaluation of gender classification across multiple head-related impulse response (HRIRs). For each angle, E0A0, E0A90, E0A180, E0A270, and E0A315, the confusion matrix, precision–recall curves, and ROC curves are presented for both female and male classes.

On this independent domain, the classification ability of all elevation azimuths ranged from 71.22% to 88.7%, suggesting that learning was stable and generalizable.

Table 1: Classification performance of the proposed CNN–BiLSTM model for gender identification using spectrograms.

Elevation	Azimuth	Accuracy	Precision	Recall	F1-score	MCC	ROC-AUC
0	0	0.9750	0.9762	0.9750	0.9750	0.9512	0.9975
0	45	0.9500	0.9500	0.9500	0.9500	0.9000	0.9975
0	90	0.8250	0.8325	0.8250	0.8240	0.6574	0.9425
0	135	0.9750	0.9762	0.9750	0.9750	0.9512	0.9975
0	180	0.9250	0.9261	0.9250	0.9250	0.8511	0.9875
0	225	0.9000	0.9000	0.9000	0.9000	0.8000	0.9850
0	270	0.8750	0.8759	0.8750	0.8749	0.7509	0.9400
0	315	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0	355	0.9750	0.9762	0.9750	0.9750	0.9512	1.0000

To further ensure gender-dependent information from identity-related cues at the signal level, independent statistical analysis was carried out with spectral and cepstral distance measures and the results are shown in **FIGURE 8**. Shapiro-Wilk analysis showed non-normality across female distributions for both feature types ($p < 0.001$), so non-parametric Mann-Whitney U tests were performed. Results showed extremely significant differences between male and female groups for spectral distance ($U = 20736$, $p = 9.92 \times 10^{49}$) and cepstral distance ($U = 3022$, $p = 2.63 \times 10^{25}$). The corresponding rank-biserial correlation coefficients further confirmed a complete distributional separation of spectral distance ($r = -1.0$) and a very large effect size for cepstral distance ($r = 0.71$). Descriptive statistics also confirmed these results: males showed significantly larger spectral distances in comparison to females, while females had consistently higher cepstral distances.

Thus it is clear that both the strategies combined with statistical tests validate our hypothesis that the HRIR encode gender specific cues and the same can be effectively derived using machine learning.

5. CONCLUSION AND FUTURE SCOPE

The results of this study using two strategies demonstrate that the HRTF-based biometric framework has the ability of reproducibly extracting and classifying gender-specific acoustic signatures using the inherent spectral temporal cues. The first strategy using spectrographic, cepstral, and deep learning analyses prove that tailored HRTFs reflect reliable, discriminative patterns from cranial geometry, pinna structure, and torso morphology, thus allowing the CNN-BiLSTM model to higher classification accuracy. For further investigating the capabilities of the proposed model, second strategy using LOSO was employed providing the accuracy in the range from 71.22% to 88.7%, verifying stable subject-independent generalization. Statistical analyses also showed an extremely significant difference in gender separability according to spectral and cepstral distances. These results confirm HRTFs as a non-trivial, technically sound, non-contact, anatomically interpretable modality of gender identification and increasing its utility in secure authentication. The research will have its future applications in personalized spatial audio and AR/VR immersive auditory interfaces.

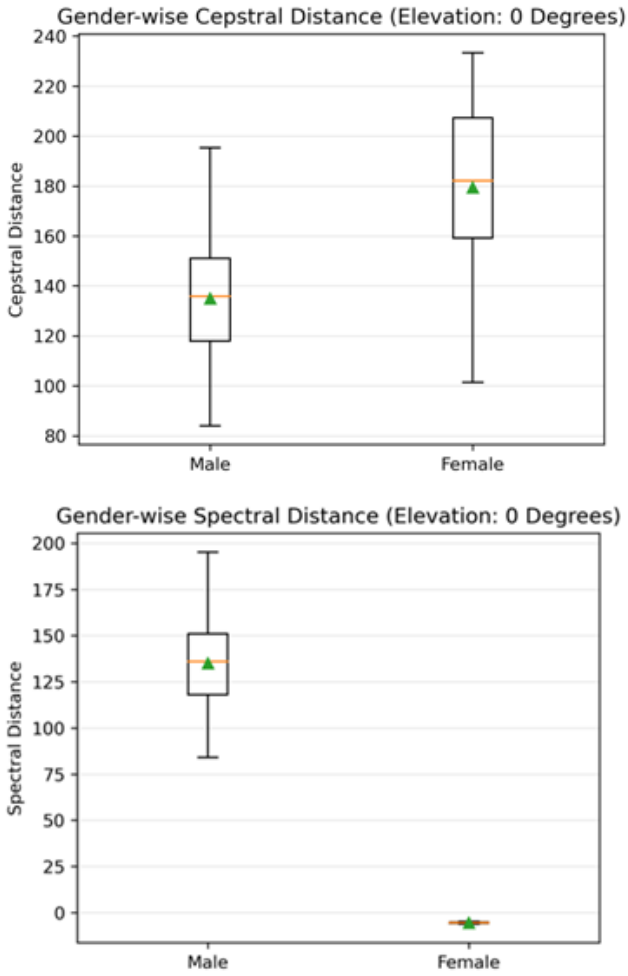

Figure 8: Gender-dependent variations in cepstral and spectral distances at frontal elevation using Mann-Whitney U test.

6. ACKNOWLEDGMENT

The authors would like to express our sincere gratitude to the University of Jammu especially the department of Electronics for providing the research facilities to complete the said research.

References

[1] Rahman AU, Halim Z. Identifying Dominant Emotional State Using Handwriting and Drawing Samples by Fusing Features. Appl Intell. 2023;53:2798-2814.

[2] Ghanbari S, Ashtyani ZP, Masouleh MT. User Identification Based on Hand Geometrical Biometrics Using Media-Pipe. In 2022 30th International Conference on Electrical Engineering

(ICEE). IEEE. 2022:373-378.

[3] Filipi Gonçalves dos Santos C, Oliveira DDS, A Passos L, Gonçalves Pires R, Felipe Silva Santos D, et al. Gait Recognition Based on Deep Learning: A Survey. ACM Comput Surv.2022;55:1-34.

[4] Vasanthi M, Seetharaman K. Facial Image Recognition for Biometric Authentication Systems Using Geometrical and Low-Level Visual Features. J King Saud Univ Comput Inf Sci. 2022;34:4109-4121.

[5] Li S, Peissig J. Measurement of Head-Related Transfer Functions: A Review. Appl Sci. 2020;10:5014.

[6] Sawhney T, Sharma A, Abrol P, Lehana PK, Yadav V, et al. Fingerprint Matching for Noisy and Distorted Patterns Using a Siamese Network With ResNet50 and Multihead Attention. IEEE Access. 2025;3:88047-88064.

[7] Jain A, Hong L, Pankanti S. Biometric Identification. Commun ACM. 2000;43:90-98.

[8] Bhatele KR, Jain S, Kataria A, Jain P. The Fundamentals of Digital Forensics. In: Gupta BB, Gupta D, editors. Handbook of Research on Multimedia Cyber Security. IGI Global. 2020:165-175.

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. Adv Neural Inf Process Syst. 2017;30.

[10] https://scholar.google.com/scholar_url?url=https://www.researchsquare.com/article/rs-1822152/latest&hl=en&sa=T&oi=gsb&ct=res&cd=0&d=6831271896214362118&ei=K9dLaYTSHJaM6rQP0-TCwQU&scisig=ALhkC2RO1tWr_F8S6ElS_biGyeHn

[11] Revathi A, Sasikaladevi N, Raju N. Real-Time Implementation of Voice-Based Robust Person Authentication Using T–F Features and CNN. Multim Tools Appl. 2023;83:31587-31601.

[12] Harar P, Burget R, Dutta MK. Speech Emotion Recognition With Deep Learning. In: 2014 Proceedings of the 4th international conference signal processing and integrated networks (SPIN). IEEE. 2017:137-140.

[13] Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, et al. End-To-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network. In: 2016 IEEE International Conference on Acoustics Proceedings (ICASSP). IEEE. 2016:5200-5204.

[14] Zhong X li, Xie B sun. Head-Related Transfer Functions and Virtual Auditory Display. Soundscape Semiotics - Localisation and Categorisation. InTech. 2014.

[15] Nagel S, Jax P. Evaluation of HRTF Models for Binaural Cue Adaptation. In: Speech Communication. 15th ITG Conference. IEEE. 2023:166-170.

[16] Brinkmann F, Lindau A, Weinzierl S. On the Authenticity of Individual Dynamic Binaural Synthesis. J Acoust Soc Am. 2017;142:1784-1795.

[17] Zekveld AA, Rudner M, Kramer SE, Lyzenga J, Rönnberg J. Cognitive Processing Load During Listening Is Reduced More by Decreasing Voice Similarity Than by Increasing Spatial Separation Between Target and Masker Speech. Front Neurosci. 2014;8:88.

[18] Schutte M. Aspects of Room Acoustics, Vision and Motion in the Human Auditory Perception of Space. [Doctoral dissertation], LMU, 2021. Available at: `https://edoc.ub.uni-muenchen.de/28543/1/Schutte_Michael.pdf`

[19] Jenny C, Reuter C. Can I Trust My Ears in VR? Literature Review of Head-Related Transfer Functions and Valuation Methods With Descriptive Attributes in Virtual Reality. Int J Virtual Real. 2021;21:29-43.

[20] Algazi VR, Duda RO, Thompson DM, Avendano C. The CIPIC HRTF Database. Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575). IEEE. 2001:99-102.

[21] Bruschi V, Grossi L, Dourou NA, Quattrini A, Vancheri A, et al. A Review on Head-Related Transfer Function Generation for Spatial Audio. Appl Sci. 2024;14:11242.

[22] Iida K. Head-Related Transfer Function and Acoustic Virtual Reality. Springer. 2019.

[23] Marschall M, Bolaños JG, Prepeli□ă ST, Pulkki V. A Database of Near-Field Head-Related Transfer Functions Based on Laser Spark Source Measurements. Appl Acoust. 2023;203:109173.

[24] Lee JW, Lee S, Lee K. Global HRTF Interpolation via Learned Affine Transformation of Hyper-Conditioned Features. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2023:1-5.

[25] Lee GW, Kim HK. Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear. Appl Sci. 2018;8:2180.

[26] Blauert J. Spatial hearing: the psychophysics of human sound localization. MIT Press. 1997.

[27] Møller H. Fundamentals of Binaural Technology. Appl Acoust. 1992;36:171-218.

[28] Xie BS. Head Related Transfer Function and Virtual Auditory Display. J. Ross Publishing. 2013.

[29] Xiang N, Schroeder MR. Reciprocal Maximum-Length Sequence Pairs for Acoustical Dual Source Measurements. J Acoust Soc Am. 2003;113:2754-2761.

[30] Yu GZ, Liu Y, Xie BS, Zhong Q. Fast Measurement System and Super High Directional Resolution Head-Related Transfer Function Database. J Acoust Soc Am. 2012;131:3304.

[31] Wightman FL, Kistler DJ. Measurement and Validation of Human HRTFs for Use in Hearing Research. Acta Acust U Acust. 2005;91:429-439.

[32] Blauert J, Brueggen M, Bronkhorst AW, Drullman R, Reynaud G, et al. The AUDIS Catalog of Human HRTFs. J Acoust Soc Am. 1998;103:3082.

[33] http://recherche.ircam.fr/equipes/salles/listen/

[34] Tzirakis P, Nguyen A, Zafeiriou S, Schuller BW. Speech Emotion Recognition Using Semantic Information. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2021:6279-6283.

[35] https://vbn.aau.dk/ws/files/227876134/1999_Minnaar_et_al_AES_Convention.pdf

[36] https://vbn.aau.dk/ws/portalfiles/portal/227975671/2000_Plogsties_et_al_AES_Paris.pdf

[37] Gardner WG. Reverberation Algorithms. In Applications of digital signal processing to audio and acoustics.Springer. 1998:85-131.

[38] Sandvad, J. Dynamic Aspects of Auditory Virtual Environments. J Audio Eng Soc. 1996.

[39] Bronkhorst AW. Localization of Real and Virtual Sound Sources. J Acoust Soc Am. 1995;98:2542-2553.

[40] Brungart DS, Kordik AJ, Simpson BD. Effects of Headtracker Latency in Virtual Audio Displays. J Aud Eng Soc. 2006;54:32-44.

[41] https://humanfactors.arc.nasa.gov/publications/Wenzel_2000_software_spatial_hearing.pdf

[42] Miller JD, Wenzel EM. Recent Developments in Slab: A Software-Based System for Interactive Spatial Sound Synthesis. In: Proceedings of the international conference Auditory Display. Kyoto. 2002:403-408.

[43] https://www.riec.tohoku.ac.jp/pub/hrtf/