Enhancing Forensic Audio Transcription with Neural Network-Based Speaker Diarization and Gender Classification

Rahmat Ullah
School of Computer Science
and Electronic Engineering
University of Essex
Colchester, CO4 3SQ, UK
Email: rahmat.ullah@essex.ac.uk

Ikram Asghar
School of Computing, Engineering
and Digital Technologies
Teesside University
Middlesbrough, TS1 3BX, UK
Email: i.asghar@tees.ac.uk

Hassan Malik
School of Computing Sciences
University of East Anglia
Norwich, NR4 7TJ, UK
Email: hassan.malik@uea.ac.uk

Gareth Evans
Faculty of Computing
Engineering and Science
University of South Wales
Pontypridd, CF37 1DL, UK
Email: gareth.evans@southwales.ac.uk

Jawad Ahmad
Cybersecurity Center
Prince Mohammad Bin Fahd University
Al Khobar, Saudi Arabia
Email: jahmad@pmu.edu.sa

Dorothy Anne Roberts
Posib Ltd
Y Gilfach, Ffordd y Pentre
Nercwys, Flintshire, CH7 4EL, UK
Email: enquiries@posib.co.uk

***Abstract*—Forensic audio transcription is often compromised by low-quality recordings, where indistinct speech can hinder the accuracy of conventional Automatic Speech Recognition (ASR) systems. This study addresses this limitation by developing a machine learning-based approach to improve speaker diarization, a process critical for distinguishing between speakers in sensitive audio data. Previous research highlights the inadequacy of traditional ASR in forensic settings, particularly where audio quality is poor and speaker overlap is common. This paper presents a neural network specifically d esigned f or g ender classification, using 20 key acoustic features extracted from real forensic audio data. The model architecture includes input, hidden, and output layers tailored to differentiate male and female voices, with dropout regularization to prevent overfitting and hyperparameter optimization ensuring robust generalization across test data. The neural network achieved an average recall of 86.81%, F1 score of 85.67%, precision of 87.95%, and accuracy of 86.83% across varied audio conditions. This model significantly improves transcription accuracy, reducing errors in legal contexts and supporting judicial processes with more reliable, interpretable evidence from sensitive audio data.**

***Index Terms*—Forensic linguistics, speech diarization, Speech transcription, Automatic speech recognition, ML**

I. Introduction

Automatic speech recognition (ASR) has witnessed significant advancements over the years, evolving from simple digit recognition systems to sophisticated models capable of transcribing complex speech in real-time. Different factors such as background noise and speaker variability pose significant challenges to ASR performance [1]. In Forensic ASR, audio presents unique challenges, characterized by poor quality and indistinct speech. Research in this domain has explored the limitations of conventional speech recognition systems when applied to forensic-like audio. A recent study provides insights into the performance of various ASR systems in transcribing audio with forensic characteristics, highlighting the gap in accuracy compared to good-quality audio [2].

Advancements in speech diarization for transcription of sensitive data using ML have been significant and multifaceted. An end-to-end speaker-attributed ASR model utilizing the "Transcribe-to-Diarize" approach was proposed [3]. This approach can manage an unlimited number of speakers and generate accurate speaker-attributed transcriptions, significantly improving traditional diarization techniques. Similarly, a speaker diarization and a hybrid ASR system specifically designed to transcribe overlapping speech conversations was proposed [4]. This system improves transcription accuracy by using speaker-specific embedding. A hybrid optimization technique called Fractional Anticorona Whale Optimization Algorithm (FrACWOA) was proposed for speaker diarization in [5].

A Speech Differentiator with an Integrated Support Vector Machine is proposed to improve user authentication in voice biometrics by differentiating between genuine and spoofed speech [6]. The system's effectiveness is validated using the ASVspoof 2019 dataset [7], focusing on multi-label classification and utilizing various ML algorithms along with critical components like frequency matching and threshold checking. The results demonstrate that SVM is superior in user identification. An overview of several deep learning models that have been used for speaker Diarization, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and their variants, can be found in [8]. Similarly, a review of various feature extraction techniques and ML classifiers employed in speech processing and recognition activities is provided in [9].

While technological advancements offer some benefits, their application in forensic audio transcription is limited and risky. Forensic audio recordings are often plagued by high levels of background noise, distortion due to compression or reverberation, and inconsistent recording conditions. These factors severely degrade speech intelligibility and present unique challenges for ASR systems. For instance, background noise can overlap with critical speech frequencies, and distortions such as clipping or compression further obscure speech content, making accurate transcription difficult. ASR systems often make errors with poor-quality recordings, leading to unreliable transcripts in forensic contexts [10]. This can potentially have unjust legal outcomes. Additionally, enhancing unclear audio does not continually improve its intelligibility and can even mislead court interpretations [11]. Transcriptions from confidential recordings are also problematic. Different transcribers, even professionals, produce varying transcripts from the same recording, indicating a need for improved forensic transcription practices [12].

This paper presents a transcription solution that can be used for sensitive audio data, specifically for police detective interviews. Initially, the project integrated advanced Natural Language Processing (NLP) techniques into a desktop application designed for audio transcription. The first application version utilized the Google Speech-To-Text API [13]. Fig. 1 provides an overview of the comprehensive software workflow diagram for audio transcription.

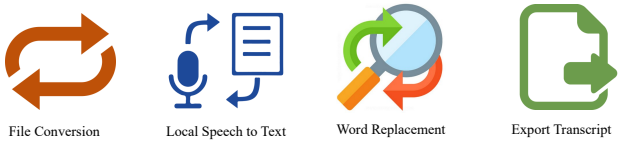

Fig. 1: Four-Stage Transcription Process: The audio files are converted to a specific format and converted to text using Google speech to Text API. Some words are replaced according to specific guidelines, and the transcription is exported and saved.

The Google Speech-to-Text model demonstrated effective performance in general transcription tasks; however, it lacked the necessary capabilities for speaker diarization, a key feature for accurate transcription in multi-speaker environments, particularly in sensitive contexts such as police interviews. To address this limitation, our research focused on developing a neural network-based system designed to classify speakers by gender, thereby enhancing speaker differentiation in transcription.

This process involved systematic data collection, cleaning, and preprocessing, followed by extensive model development and evaluation. By training the neural network on 20 key acoustic features extracted from real forensic audio, we aimed to optimize speaker identification accuracy. The primary goal of this system is to accurately determine speaker gender, which is essential to improving the precision of diarization. This enhancement is critical to ensuring the reliability and integrity of transcriptions in sensitive environments where accurate speaker identification can significantly impact the quality and admissibility of evidence.

II. Methodology

A. Data Preprocessing

The data set comprises 3,168 voice samples, representing a balanced mix of male and female voices, and can be found in [14]. These samples were collected from various sources to ensure diversity and representativeness.

The voice samples were preprocessed and 20 key acoustic features, essential for effective gender differentiation, are extracted using WarbleR package in R [15]. The features included fundamental frequency, intensity, formants, and other spectral characteristics such as spectral entropy, spectral flatness, frequency centroid, and modulation index relevant to voice and speech analysis. This resulted in a 3168x22 feature matrix (one row for each observation). Next, we perform data normalization to ensure that features with more extensive numerical ranges do not overpower those with smaller ranges during training. We used the scikit-learn library to normalize the data in the range of [0, 1] according to equation (1):

$$x_{\text{normalized}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{1}$$

Where X_{min} and X_{max} are the minimum and maximum values of the feature X, respectively.

In this study, the test data comprised specifically curated audio clips, which were used to evaluate the model's performance. These clips were carefully processed to eliminate any silence or cross-talk, ensuring the dataset was optimally prepared for testing purposes. Each audio clip was manually labeled, and the metadata was systematically compiled into a CSV file through an R script. This meticulous preparation highlights the significance of assessing the model under conditions that closely mimic its intended real-world applications.

Further evaluations were conducted using sequential 3-second audio clips that included noise, silence, and cross-talk,

introducing additional complexity to simulate challenging real-world conditions. This phase of testing was crucial for examining the model's robustness and ability to handle diverse audio characteristics. Such rigorous testing protocols are essential for identifying potential weaknesses in the model's design and its efficacy in processing unstructured data.

B. Feature Extraction

Kernel Density Estimate (KDE) plots were used to visually analyze the distribution of various features within the unseen test data, categorized into female" and male" classes. KDE plots provide a visual representation of the probability density function of continuous variables, illustrating the likelihood of a variable falling within a specific range of values, which is valuable for identifying distribution patterns within the data.

The analysis produced a series of 20 KDE plots, each corresponding to a specific feature of the data set, as shown in Fig. 2.

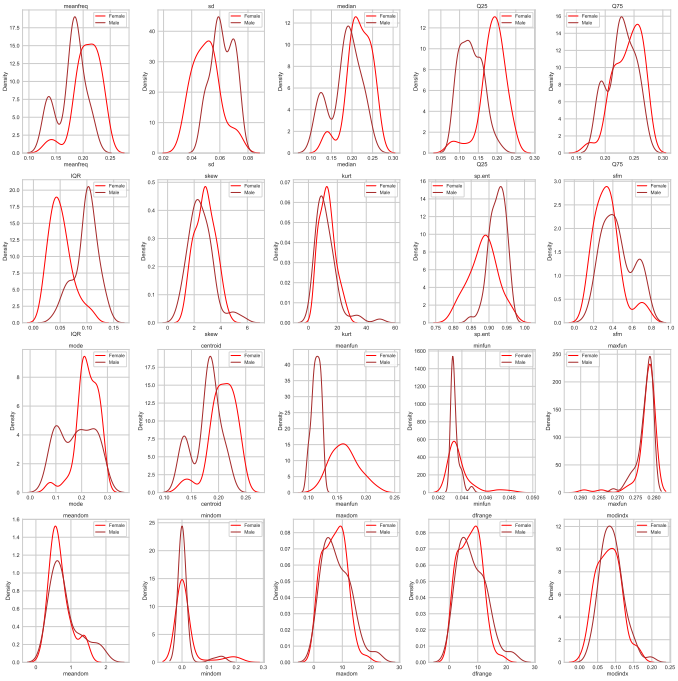

Fig. 2: Kernel Density Estimates of Acoustic Feature Distributions by Gendee. The figure showcases a set of kernel density estimates for various features, illustrating the probability density functions of each acoustic feature for female (red) and male (blue) voice samples. These smooth curves facilitate the identification of distribution patterns, such as central tendency, dispersion, and potential skewness, enabling a clear visual comparison between the two gender classes for each feature in the dataset.

These visualizations highlighted various distribution characteristics, such as multimodality, skewness, and other peculiarities in the data that could influence the performance of the classification mode. Using these plots, it is possible to assess the discriminative power of each feature for the gender classification task based on voice data. Features with significant overlap in densities between classes were identified, suggesting a lower discriminatory power and potentially influencing feature selection and preprocessing decisions. Conversely, features that exhibited distinct distributions between classes were noted for their potential usefulness in enhancing the model's ability to differentiate between female" and male" classes.

The correlation analysis of acoustic features extracted from voice data was performed to understand the interdependencies between variables and their relationship with the target variable. A correlation matrix was constructed to measure the pairwise correlations between all features and depicted in the heatmap in Fig. 3.

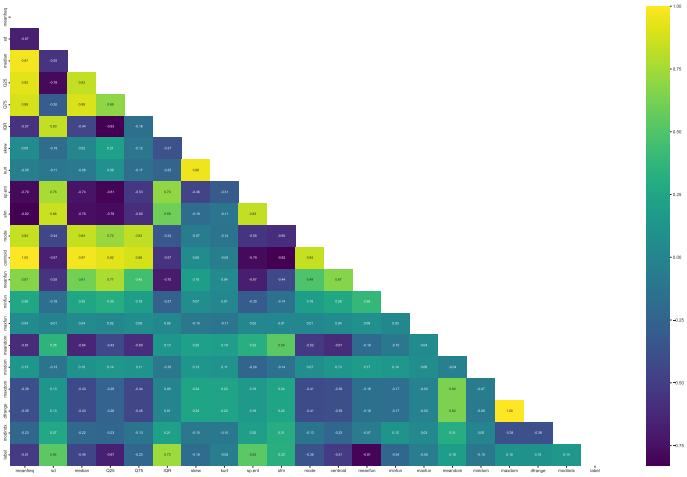

Fig. 3: Corelation matrix. The correlation coefficients between different pairs of features are depicted in off-diagonal cells. The color indicates the strength and direction of the relationship.

The matrix represents correlation coefficients from -1 to +1 with darker shades corresponding to stronger (positive/negative) correlations, while lighter shades indicate weaker correlations. This matrix is crucial in detecting multicollinearity, a phenomenon where two or more independent variables in a regression model are highly correlated. This condition can lead to increased variance in the coefficient estimates, destabilizing the model. This can lead to overfitting and difficulties in interpreting the effects of individual features.

The analysis revealed that certain features had a more significant linear relationship with the target variable than others. These features are prime candidates for inclusion in the machine learning model, as they are likely to contribute meaningfully to the model's predictive accuracy. Conversely, features with weak or negligible correlation were excluded from the model to simplify it and reduce computational complexity.

C. Neural Network Implementation

We trained a neural network exclusively on the dataset described above for the task at hand. This dataset does not

contain any sequential 3-second data that would be used for subsequent testing. The developed neural network was explicitly tailored for speaker diarization based on the classification of gender from voice samples. The architecture was intentionally designed to be straightforward, avoiding unnecessary complexity to enhance understandability and efficiency. It featured an input layer, a hidden layer, and an output layer with two classes, male and female. Dropout was used by randomly disabling neurons to prevent the model from overfitting, thus promoting the generalization capability of the model.

The model's efficacy depends significantly on optimizing a set of hyperparameters, which provides extensive opportunities for experimental refinement. These hyperparameters include the train/test split ratio, the number of K-fold in cross-validation splits, random state settings, the number of neurons in each layer, dropout rates, activation functions, loss functions, optimizer choices, batch sizes, learning rates, and the total number of training epochs.

For model training, Keras offers a suite of hyperparameter optimization algorithms that adjust the weights and biases within the network to minimize loss functions effectively. We selected Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam) for their robust performance in various test scenarios. The choice of optimizer is pivotal, as each algorithm has unique characteristics suited to different types of data and network structures. Regarding loss functions, which quantify the gap between the predicted outputs and actual labels, we incorporated Mean Squared Error (MSE), Mean Absolute Error (MAE), and Binary Cross-Entropy (BCE). These functions are instrumental in guiding the training process by providing a measurable indicator of prediction accuracy, which is vital for adjusting model parameters effectively.

The robustness of our model was evaluated using 10-fold cross-validation. This method enhances the reliability of the performance assessment by ensuring that each subset of the data is used for both training and validation. After training, the model's performance is assessed using metrics calculated from each fold's test data, with stringent controls to prevent any data leakage.

Furthermore, to evaluate the overall effectiveness of our classifier, we generated Receiver Operating Characteristic (ROC) curves and Precision-Recall curves using the test datasets. These curves are invaluable for examining the trade-offs between sensitivity (True Positive Rate) and specificity (False Positive Rate), crucial metrics for binary classification tasks. The True Positive Rate (TPR) and False Positive Rate (FPR) are calculated using the equations:

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

The TPR indicates how effectively the classifier detects positive instances. In our model, a high TPR indicates its capability to correctly identify positive cases, while a low TPR suggests inefficiency in this region. Similarly, the False Positive Rate (FPR) assesses the proportion of actual negative cases that our binary classifier incorrectly identifies as positive. It is calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

The FPR is used alongside the TPR to generate the ROC curve, which comprehensively evaluates the model's performance. These metrics provide insights into the classifier's ability to detect positive instances accurately while minimizing false positives, thus offering a comprehensive view of model performance. In our neural network model, a low FPR indicates proficiency in avoiding false alarms, whereas a high FPR suggests a higher propensity for false alarms. Our approach to model training and validation not only enhances the prediction accuracy but also ensures that the neural network can generalize well to new, unseen data, thereby validating the efficacy of our methodology in real-world applications.

III. Results and Analysis

The neural network model demonstrated remarkable performance in gender classification, achieving a training and validation accuracy of 98 %, and a test accuracy of 94.4 %. These metrics underscore the model's capability to generalize effectively to unseen data, a crucial aspect of practical machine learning applications.

Precision, recall, and F1-score were closely monitored, reflecting high values that confirm the model's balanced capacity for sensitivity and specificity. The precision, indicating the proportion of positive identifications that were actually correct, and the recall, indicating the proportion of actual positives that were correctly identified, led to a robust F1-score. These metrics highlight the model's effectiveness, optimizing it for real-world deployment where the accuracy of classification is critical.

The Area Under the ROC Curve (AUC) is employed as a standard metric to evaluate the model's performance. The AUC value ranges between 0 and 1, with 1 indicating a perfect classifier and 0.5 representing a random classifier. A higher AUC value signifies superior overall performance. The ROC curve visually illustrates the trade-off between sensitivity and specificity, which is crucial in selecting an appropriate threshold depending on the desired balance between the two. The results of our trained model are presented in Figure 4.

As anticipated, the neural network model's high training accuracy translated into exceptional precision and recall scores for the validation set, yielding a remarkable F1 score. The model's average accuracy stands at an impressive 98.63%. These results not only validate the effectiveness of the neural network architecture and training regimen but also highlight the potential for its application in more complex scenarios requiring nuanced differentiation between classes.

A. Neural Network Predictions on Unseen Test Data

The evaluation of the neural network on unseen test data provided significant insights into its predictive accuracy and generalization ability. The model's performance was assessed

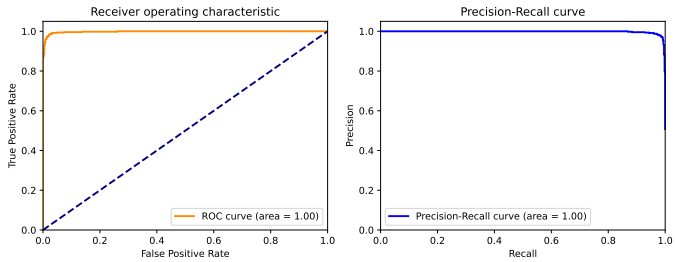

Fig. 4: Receiver Operating Characteristic (ROC) and Precision-Recall Curves demonstrating ideal performance with an area under curve (AUC) of 1.00 for both metrics, indicative of good performance.

using the ROC and Precision-Recall curves, as depicted in Figure 5.

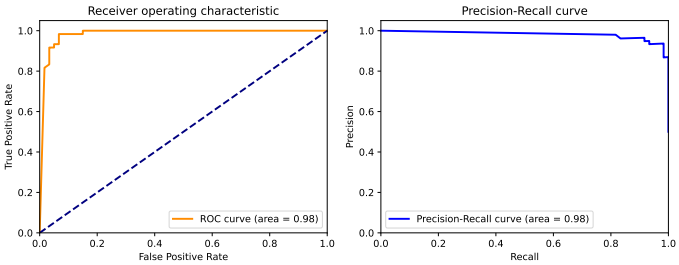

Fig. 5: Performance evaluation of the neural network model using ROC and Precision-Recall Curves for unseen data. The right panel presents the Precision-Recall curve, also with an AUC of 0.98, indicating a high level of precision across varying thresholds. These curves demonstrate the model's robust discriminative power and predictive reliability when applied to unseen test data.

The ROC and Precision-Recall curves both achieved an AUC score of 0.98, indicating exceptional separability and precision, respectively. These outcomes highlight the model's strong performance, maintaining high accuracy even when tested on new and previously unseen data, thus confirming its robustness and generalization capabilities.

B. Speaker Diarization

To explore the feasibility of speaker diarization using the trained gender classification model, tests were conducted with sequential 3-second audio data segments, representing a challenging scenario with no preprocessing steps such as noise, cross-talk, and pause removal. The dataset comprised various difficulty levels, ranging from 'Easy' to 'Hard' and was intended to evaluate the model's performance in a realistic conversation setting between male and female speakers.

The audio files were split into 3-second segments and then sorted and processed to generate data, which required manual labeling for gender classification. Each difficulty-level file was loaded into separate data frames for analysis. Before the neural network deployment for gender classification, the test data was normalized to ensure consistency with the scale of the training data. This step is critical to the predictive accuracy of the model, as it helps to eliminate any bias that varying scales among the features may introduce. The normalization process was carried out using a predefined function, which applies the""transfor"" method of the scaler object used during the training phase. This method adjusts the test data to the same scale as the training data without altering the scale parameters, thereby maintaining the model's integrity.

The normalization process was performed separately on each CSV file containing the unseen test data for different difficulty levels, labeled as 'Easy', 'Medium", and 'Hard'. The features within each dataset were normalized, and the data was then split into input features and output targets to be fed into the neural network.

The model's diarization performance was measured against these manually labeled audio segments. The feasibility of this approach in real-world applications would depend on the model's accuracy in classifying each segment as either 'Male' or 'Female'. The data and the segment labels served as the ground truth for evaluating the model's precision in distinguishing between male and female speakers within the context of a conversation.

The neural network performance on unseen sequential 3-second test data across various difficulty levels was quantified using several metrics. The summarized results are presented in Table I, highlighting the recall, F1 score, average precision, and average accuracy for each dataset.

TABLE I: Neural network performance on unseen sequential 3-second test data. Note: All values are in percent, focusing on average metrics and key scores.

Dataset	Recall	F1 Score	Avg Precision	Avg Accuracy
Easy 02	88.24	90.91	98.15	94.10
Easy 06	95.24	86.96	89.57	88.37
Medium 09	89.80	88.89	86.96	88.88
Medium 10	90.24	84.09	79.77	82.41
Hard 08	86.36	82.61	78.37	81.02
Hard 09	70.97	78.57	90.89	84.20

As seen in Table I, the 'Easy02" dataset showed exemplary performance with the highest average accuracy (94.10%). Conversely, the"Hard 0" and"Medium10" datasets exhibited lower average accuracies (81.02% and 82.41%, respectively), indicating a decline in model performance as the difficulty of the audio conditions increased. Nevertheless, the average precision remained relatively high across all datasets, demonstrating the model's consistent ability to identify the correct gender in mixed-gender conversations. The results suggest that while the model is robust under less challenging conditions, its performance is affected by the complexity of the audio data, as expected in real-world scenarios.

To evaluate the gender classification capability in speaker diarization, the actual gender labels were compared with the model's predictions for different audio segments. For this purpose, the resulting data was compiled into individual CSV files, each containing the predicted and actual labels for audio

segments. This gives us a granular view of the model's diarization capabilities. It allows for an in-depth analysis of how well the model can identify and differentiate between speaker's genders in a sequence of audio clips. This is particularly relevant for sensitive data, where accurate gender identification is crucial.

IV. Conclusion and Future work

This study evaluated the effectiveness of neural network models for speaker diarization, specifically focusing on gender classification within dialogues. The model achieved an impressive 86.5% accuracy on sequential 3-second audio clips without advanced preprocessing, demonstrating the substantial potential for audio transcription applications. These results highlight the critical influence of data quality and speaker expressiveness on the model's performance. Differences in accuracy across various testing scenarios underscored the model's sensitivity to audio input quality, particularly in more challenging real-world conditions. The research emphasized that meticulous data preparation is crucial, with well-prepared input data significantly enhancing model predictions. Additionally, the study identified the selection of appropriate time windows for audio processing as a critical factor affecting accuracy. Optimizing this parameter could further improve the model's effectiveness in diverse audio environments.

While the model demonstrates robust performance under controlled conditions, its effectiveness diminishes in scenarios with severe audio distortions or heavily overlapping speakers. However, the high precision and recall metrics underscore its reliability in moderately challenging conditions, making it a valuable tool for forensic applications. Future research should focus on refining data preprocessing techniques, expanding the training dataset, and exploring data augmentation to boost model robustness and adaptability. This foundational work paves the way for advanced developments in automated speaker diarization, aiming to create more accurate and reliable transcription tools.

References

[1] O. R. Khazaleh and L. A. Khrais, "An investigation into the reliability of speaker recognition schemes: analysing the impact of environmental factors utilising deep learning techniques," *Journal of Engineering and Applied Science*, vol. 71, no. 1, pp. 1–24, 2024.

[2] "Frontiers — automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare?" https://www.frontiersin.org/articles/10.3389/fcomm.2024.1281407/abstract, (Accessed on 01/26/2024).

[3] N. Kanda, X. Xiao, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8082–8086.

[4] S. R. Chetupalli and S. Ganapathy, "Speaker conditioned acoustic modeling for multi-speaker conversational asr," *arXiv preprint arXiv:2104.01882*, 2021.

[5] K. VijayKumar *et al.*, "Optimized speaker change detection approach for speaker segmentation towards speaker diarization based on deep learning," *Data & Knowledge Engineering*, vol. 144, p. 102121, 2023.

[6] M. A. Basit, C. Liu, and E. Zhao, "Sdi: A tool for speech differentiation in user identification," *Expert Systems with Applications*, vol. 243, p. 122866, 2024.

[7] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[8] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[9] K. B. Bhangale and K. Mohanaprasad, "A review on speech processing using machine learning paradigm," *International Journal of Speech Technology*, vol. 24, pp. 367–388, 2021.

[10] D. Loakes, "Does automatic speech recognition (asr) have a role in the transcription of indistinct covert recordings for forensic purposes?" *Frontiers in Communication*, vol. 7, p. 803452, 2022.

[11] H. Fraser, "'enhancing'forensic audio: what if all that really gets enhanced is the credibility of a misleading transcript?" *Australian Journal of Forensic Sciences*, vol. 52, no. 4, pp. 465–476, 2020.

[12] R. Love and D. Wright, "Specifying challenges in transcribing covert recordings: Implications for forensic transcription," *Frontiers in Communication*, vol. 6, p. 797448, 2021.

[13] "Speech-to-Text: Automatic Speech Recognition — Google Cloud," https://cloud.google.com/speech-to-text, online; accessed 26 January 2024.

[14] K. Becker, "Gender Recognition by Voice," https://www.kaggle.com/datasets/primaryobjects/voicegender, online; accessed 1 Feb 2024.

[15] M. Araya-Salas and G. Smith-Vidaurre, "warbler: an r package to streamline analysis of animal acoustic signals," *Methods in Ecology and Evolution*, vol. 8, pp. 184–191, 2017.