VOICE IDENTIFICATION AND SPEECH RECOGNITION: AN ARENA OF VOICE ACOUSTICS

Sarita Sheoran[1], Dr. Dolly Mahna[2]*

Abstract

Forensic voice comparison (i.e., forensic speaker recognition, identification, and voice comparison) is a subdiscipline of forensic science in determining the authenticity of questioned voices by comparing the analytical results and drawing inferences from the comparison of reference and suspect voice recording. The comparison provides an overview of analytical accessions and interpretative structures that were used in legal admissibility and validation. Various aural-perceptual acoustic and phonetic features such as fundamental frequency(F0), vowel formants, spectral characteristics, dialect, voice quality, articulation, stress pattern, intonation, nasality, prosody, non-fluencies, speech disorders, and diseases are taken into consideration for speaker profiling. This review focuses on various physical, psychological, and mechanical aspects of vocalization, acoustic recognition, analysis, and comparison that involves pre-processing, various feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Discrete Wavelet Transform (DWT), and Formants Wavelet Entropy (FEW) along with feature matching (technique) (that) utilizing GMM(Gaussian Mixture Models), SVM(Support Vector Machines), DNN's(Deep Neural Networks) algorithm along with the AuPhA (Auditory Phonetic and Acoustic) approach. The Likelihood Ratio (LR)and ANOVA (Analysis of Variance) methods are used based on the extraction and analysis using voice parameters such as pitch, formants, and spectrogram for investigating and finding the impersonator/culprit.

Keywords: Voice comparison, speaker profiling, acoustic parameters, forensic science, Extraction, vocalization.

[1]M.Sc. Student, Department of Forensic Science, Chandigarh University, Punjab, India
Email: misusingh25@gmail.com
[2]*Assistant Professor, Department of Forensic Science, Chandigarh University, Punjab, India
Email: dolly.e13936@cumail.in

***Corresponding Author:** Dr. Dolly Mahna
*Assistant Professor, Department of Forensic Science, Chandigarh University, Punjab, India
Email: dolly.e13936@cumail.in

DOI: - 10.31838/ecb/2023.12.si5.008

Introduction

With a constantly increasing demand for voice forensics worldwide, there has come a progressively stringent necessity for acceptance of forensic scientific evidence in the court of law, as reproducible, impartial, logical factuality, reliable, and demonstrated rationality, etc. (Morrison G. S., 2014). Forensic voice analysis is an examination conducted for comparison of questioned voice recordings with the known voice about their phonetics and acoustics and helps the judiciary figure out the identity of the impersonator/ doubted speaker (Singh, 2017; Enzinger, 2017)

As a discipline of Forensics, Forensic phonetics, and forensic voice comparison tool (as well) undergoes a remarkable paradigm deviation/shift in the examination/analysis and assessment of Forensic evidence. (Wang H. &., 2020)

The relevant areas of Forensic voice are: -

i. Forensic Phonetics aims at the examination of spoken communication, it intends to enhance and decode a spoken message, analysis of emotions in speech, and speaker identification along with deciding the legitimacy of voice recordings (Jessen, 2008).
ii. Forensic Linguistics which includes psycholin guistics points out language that is analyzed to regulate origination, deception, and the intention of an individual, etc., (Schilling, 2015) This also includes speech decoding.
iii. Forensic Psychoacoustics comprehend auditi-on and human aural faculty, which involves heard signs and their acoustic, neural, and perceptual effects on an individual and their behavior (Hollien H. B., 2014).

With the advancement of technology, modes of communication are changing rapidly in the technological era. The mobile phone sets are easily accessible and affordable means with inbuilt multimedia features and various other revolutionary digital recorders with small-size memory chips that are available and are easy to handle and access (Goyal, 2019). Investigating agencies use the well-established method of identification through voice., if there is a recorded conversation as physical evidence (Goyal, 2019).

Acoustic voice analysis based on disturbances, and measures have been the objective for a long debate considering its validity and fundamentally its efficacy of criteria for perceptual evaluation, a standard, and point of reference for examining voice quality (Batalla, 2014). Technically termed as forensic speaker identification, in cases when the suspected individual refuses/is not available to give the specimen voice sample (Morrison G. S., 2016), then it's difficult for the forensic labs or agencies to prove the involvement of that individual in any criminal activity as speaker identification is not feasible in the absence of control samples, Speaker profiling aids the investigating labs/ personals to identify the true criminal and to narrow down the investigation (Morrison G. S., 2016). Speaker Profiling consists of the extraction of personal information concerning the anonymous offender from implicated speech information/matter such as age, gender, height, vocalization, dialect, features of respiration, articulation, and way of speaking (Hughes V. &., 2015; Hansen, 2015; Albuquerque L. O.-C., 2020). So, an individual belonging to a particular regional dialectal group/ region could be recognized with acoustic attributes reproduced in his/her oration.

In voice recognition particular /specific details gathered from oration could be directed either automatically or manually. Manual recognition is intuitive, subjective, and susceptible to hearing and acknowledgment of human standards (Todkar, 2018). Automatization directs objectivity and speed of recognition. Voice identification could be applied as discourse recognition I.e., identification of speech or phonetics of signals, and speaker recognition i.e., to identify a speaker (Jessen, 2008; Todkar, 2018). Even after a different approach, both speaker and discourse recognition face the same plight. Noise, phonetics, and accent are a few elements that make pattern identification instruments applied to voices challenging (Hughes V. &., 2020).

Within the arena of speaker identification, it divides the discipline into recognition and verification (Kaur, 2015). Speaker verification proceeds when the system collates a confronted discussion with reference discourse such that it substantiates if one or the other was produced by the very same individual. Speaker identification corresponds to the behavioral and physiological features of individual articulation (De Lara, 2018; Khelif, 2017). And speaker identification works on various statistical techniques for model speakers as well as to compensate for network /session discrepancies in oration data. Typically, speaker recognition uses MFCC (Mel Frequency Cepstral Coefficients) characteristics for specifications of speech and utilized either discriminative or generative models for pattern classification (Xian Y. M., 2013; Gerlach, 2020; Walsh, 2007). GMM (Gaussian Mixture Models) along with maximum a posterior (MAP) adapted GMM universal background model

(GMM-UBM) falls under the generative modeling group (Morrison G. S., 2016), while SVM (Support vector machines) belongs to the discriminative model (Haris, 2015). Currently, the major speaker recognition system utilizes the low dimensional representation of GMM super vectors called the i-vectors, derived from factor analysis for describing speaker utterances. GMM and i-vector methods with PLDA (Probabilistic Linear Discriminant Analysis) are compared for Text Independent Speaker Recognition (SR), and Neural network (NN) (Choi, 2020) (Eskidere, 2016). The two types of features are PNCC (Power Normalized Cepstral Coefficients) and RASTA PLP – Relative Spectral Perceptual Linear Prediction Coefficients and Vector Quantization (VQ) (Segundo E. S., 2019). It's presented that PNCC features have a closer approximation of human aural faculty than RASTA-PLP (Nayana, 2017; Univaso, 2017).

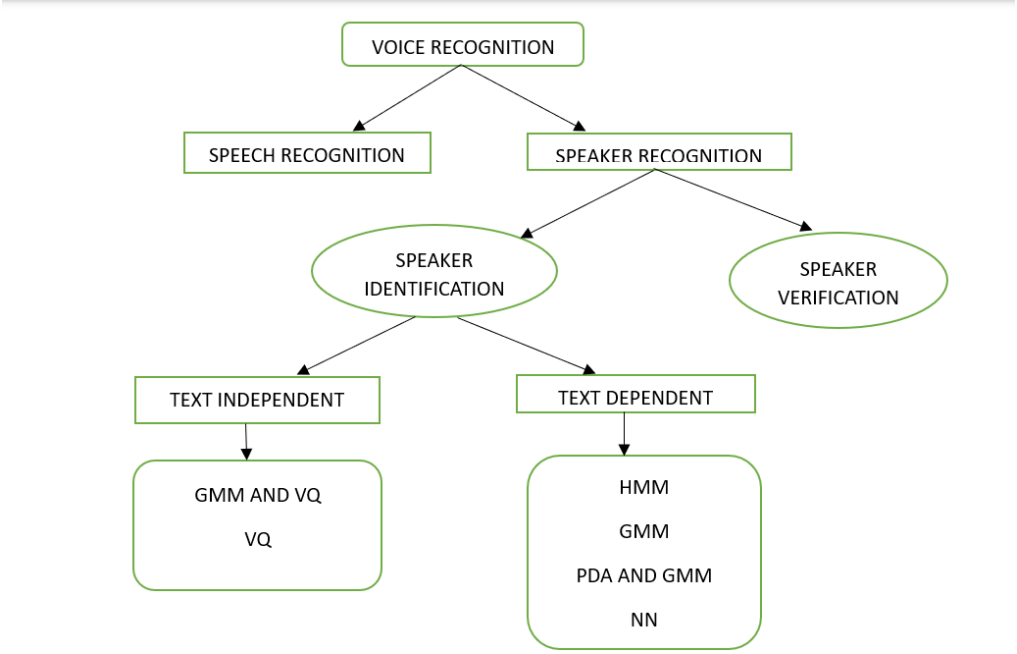

Figure 1:- Voice Analysis Classification

Table 1:- Types of speaker recognition and their effects.

Method	Description	Advantage	Disadvantage
Text Dependent (Todkar, 2018; Brown, 2017; Kim, 2021)	Requires a specific phrase or set of phrases to be spoken	High accuracy, resistant to impersonation attacks	Limited flexibility requires pre-registration of phrases and speaker
Text Independent (Todkar, 2018; Ahuja P. &., 2018)	Does not require a specific phrase to be spoken	High flexibility, no pre-registration required	Lower accuracy than text-dependent systems, vulnerable to impersonation attacks
Speaker verification (Todkar, 2018)	Verifies the identity of a speaker based on their voice characteristics	Non-invasive can be used for real-time verification, resistant to spoofing attacks	Accuracy can be affected by changes in voice due to illness, age, emotional state, and other factors; requires a known sample of the speaker's voice for comparison
Speaker identification (Todkar, 2018)	Determines the identity of speakers based on their voice characteristics	Can be used to identify unknown speakers, and can provide strong evidence in forensic investigations.	Requires a large database of known speakers, accuracy can be affected by environmental factors and variations in speech

Human acoustics have detailed phonetic cues that contain fundamental frequency(f0), intensity, duration, vocal tract spectrum, and spectral tilt has the ability to perceive emotions like happiness, anxiety, fear, anger, and deception, (Gamer, 2006) depression, fatigue, etc., (Xian Y. M., 2016; Amin, 2014) Fundamental frequency decides the emotional status of the speaker and is the frequency of vibration i.e., the opening and closing of vocal cords per second (Sondhi, 2015).)

Formants show the spectral arrangement as a function of time and a single formant correlate to the natural vibration of the vocal tract (Xian Y. M., Reliability of human-supervised formant-trajectory measurement for forensic voice comparison., 2013; Cenceschi, 2021). The spectral peaks i.e., formants are obtained when these effects of resonances speculate in the sound spectrum and they amount to the concentration of acoustic energy throughout a specific frequency in the speech and are produced

in approximately 1000Hz intervals/ one in each 1000Hz band. It has been observed that psycho-physiological stress along with the fundamental frequency of voice diverges from its baseline.

Mechanism of voice production

Organs of speech are the active vocal organs that are involved directly or indirectly in the production of speech (Kulshreshtha, 2012). The vocal tract comprises of lungs, chest muscles, trachea, larynx, pharynx, lips, teeth, tongue, the roof of the mouth, palatine, and vocal folds. When someone intends to produce speech/voice, the brain first sends the set of signals to the organs responsible for speech production. (Madill, 2020)Sound is produced in the vocal tract when the air is expelled out through the vocal cords, causing them to vibrate (Kulshreshtha, 2012). The vibration of vocal cords creates sound waves that travel up through the throat and into the mouth and nasal cavity when they are modified by the shape and movement of the vocal tract that produce different sounds and tones. (Madill, 2020)The vocal cords themselves are located in the larynx/voice box which is situated at the top of the windpipe. The cords are made up of two folds of tissues that are stretched across the opening of the larynx. When air from the lungs passes through the cords, they vibrate rapidly, producing sound (Brockmann-Bauser, 2023; Arabi, 2023).

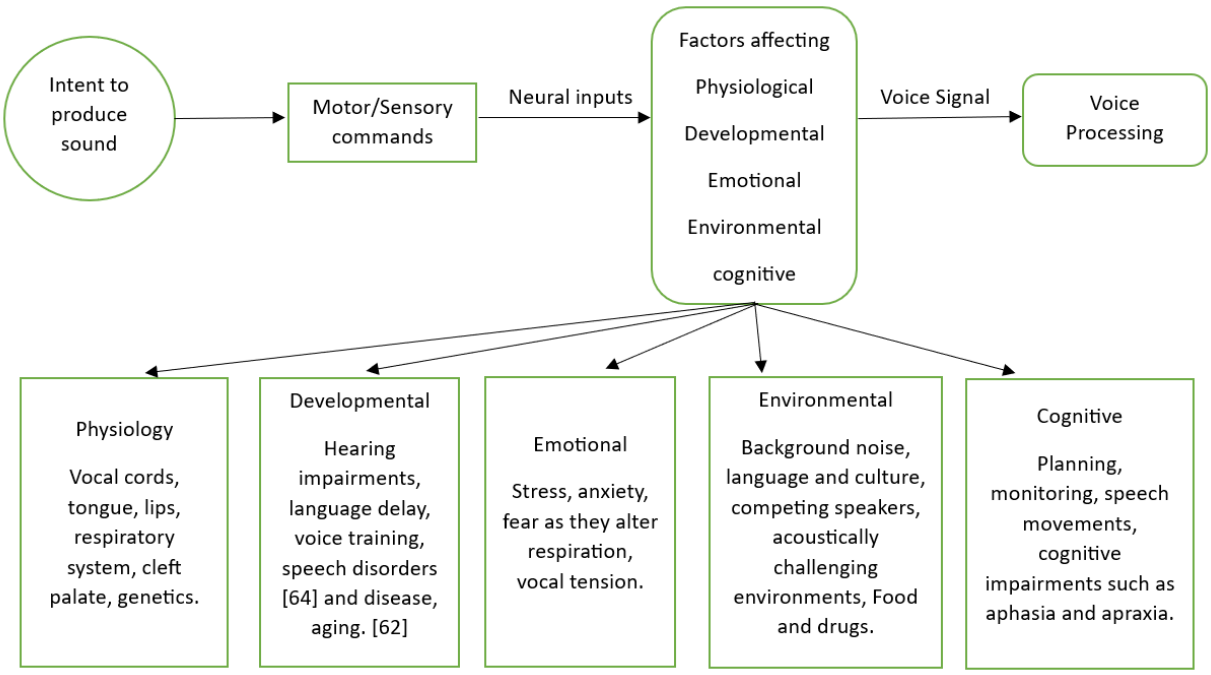

Figure 2:- Mechanism and factors affecting voice production.

Supra Segmental features

Basic components of speech, that form an utterance in conjunction with vowels and consonants segment known as syllable, these features are also referred to as prosodic features (Ahuja P. &., 2018). Suprasegmental features include stress, intonation, tempo, pitch, length of segment, and rhythm (Ahuja P. &., 2018; Kulshreshtha, 2012).

Pitch- refers to the highness or lowness of a speaker's voice (Hollien H. B., 2014). Fluctuations in pitch can convey differences in meaning or tone (Sondhi, 2015).)

Stress- refers to the emphasis placed on syllables within a word. Stressed syllables are typically pronounced louder, longer, and with a higher pitch than an unstressed syllable. Stress distinguishes between words or conveys emphasis (Harnsberger, 2009; Ahuja P. &., 2018).

Intonation- refers to the rise and fall of pitch across a sentence or utterance. variations in intonation convey the differences in meaning or attitude (Ahuja P. &., 2018).

Tempo- Refers to the rate at which a speaker speaks. Tempo changes change the differences in meaning or emotion. (Ahuja P. &., 2018)

Length of the segment- it's the measure of the duration of the vowel used in the segment. It conveys the differences in meaning with different vowel lengths (Ahuja P. &., 2018).

Rhythm- refers to the pattern of stressed and unstressed syllables in voice, it usually affects the flow and pace of speech (Hollien H. B., 2014).

Acoustic voice quality parameters

Acoustic voice quality parameters are measures used to describe the physical characteristics of a

person's voice (Batalla, 2014). These parameters are often used in voice analysis and can provide information about a person's vocal health, age, vocal habits, and overall voice quality (Albuquerque L. O.-C., 2020). These parameters can provide information about the pitch, loudness, and timbre of the voice. Some common acoustic voice quality parameters include:

Fundamental frequency (F0): This is the rate at which the vocal folds vibrate and is perceived as pitch. It is measured in hertz (Hz) (Batalla, 2014; Sondhi, 2015).).

Intensity: This refers to the loudness of the voice and is measured in decibels (dB) (Hollien H. B., 2014).

Jitter: This parameter measures the cycle-to-cycle variation in the fundamental frequency and is often used as an indicator of voice instability (Batalla, 2014). It is measured as a percentage. (Grillo, (2020)

Shimmer: This parameter measures the cycle-to-cycle variation in the intensity of the voice and is often used as an indicator of voice quality (Batalla, 2014). It is measured as a percentage. (Grillo, (2020)

Harmonics-to-Noise Ratio (HNR): This is the ratio of energy in the harmonic components of the voice to the energy in the noise components (Batalla, 2014). A higher HNR is associated with clearer and more pleasant voice quality. (Grillo, (2020) (Madill, 2020).

Formant frequencies: These are resonant frequencies of the vocal tract that contribute to the timbre of the voice (Grillo, (2020). They are typically represented as the first and second formant frequencies (F1 and F2) (Batalla, 2014).

The voice breaks: This refers to sudden changes in the voice quality due to the instability of the vocal folds (Wang Q. Z., 2021c).

Spectral balance: This refers to the distribution of energy across different frequency bands in the voice sound (Brockmann-Bauser, 2023).

Vibrato: This refers to the small variations in pitch, loudness, or timbre that occur naturally in the human voice (Grillo, (2020).

Spectral tilt: This refers to the slope of the spectrum of the voice sound, which can indicate the size and shape of the vocal tract (Brockmann-Bauser, 2023).

Factors affecting speech recognition accuracy and performance

Background noise: Ambient noise, such as traffic, crowd noise, or wind, can interfere with the clarity of speech and reduce the accuracy of speech recognition systems. Noise reduction techniques, such as spectral subtraction or Wiener filtering, can be used to mitigate the effect of noise (Denk, 2014).

Speaker variability: Differences in voice quality, accent, pronunciation, and speech rate among speakers can affect speech recognition accuracy. Speaker adaptation techniques, such as speaker normalization or feature mapping, can be used to adjust for these differences (Brockmann-Bauser, 2023).

Speech variability: Variations in speech content, context, and style can affect speech recognition performance. Techniques such as language modeling, which uses statistical models to predict likely sequences of words, can be used to improve recognition accuracy (Xian Y. M., 2013).

Channel variability: Variations in the recording environment, microphone quality, or transmission channel can affect the quality of the speech signal and reduce recognition accuracy. Channel compensation techniques, such as equalization or de-reverberation, can be used to enhance the speech signal (Wang Q. Z., 2021c).

Vocabulary size: The size and complexity of the vocabulary can affect recognition performance. Larger vocabularies require complicated prototypes and may escalate the possibility of recognition errors (Morrison G. S., 2023).

Training data: The quality, quantity, and diversity of the training data used to train the speech recognition system can affect its accuracy and robustness. A larger and more diverse training dataset can improve system performance (Wang Q. Z., 2021c).

Speaker Recognition Process

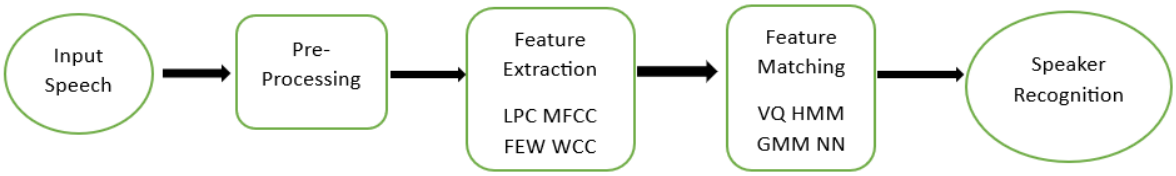

Figure 3:- Speaker Recognition process and implementation.

Pre-Processing

Pre-processing is an important step in speaker recognition systems to improve the quality of speech signals and to extract relevant features that can be used for speaker identification, it helps to enhance the quality of the audio and reduce noise and other distortions (Todkar, 2018).

Feature extraction techniques

Table 2:- Various feature extraction techniques for Voice analysis.

Feature Extraction Technique	Description
Mel-Frequency Cepstral Coefficients (MFCC)	A commonly used technique that converts the speech signal into a series of spectral bands, then applies the discrete cosine transform to obtain a set of coefficients that represent the spectral envelope of the speech signal (Wang Q. Z., Towards Query-Efficient Adversarial Attacks Against Automatic Speech Recognition Systems, 2021; Kaur, 2015).
Linear Frequency Cepstral Coefficients (LFCCs) (Aljasem, 2021)	Similar to MFCCs, but uses a linear frequency scale instead of a Mel frequency scale.
Discrete Wavelet Transform (DWT)	A variant of the wavelet transform that uses a specific set of wavelets (Daubechies, Coiflet, etc.) to decompose the signal into coefficients
Linear Predictive Coding (LPC)	A technique that models the spectral envelope of a speech signal using a linear filter. The filter coefficients are used as features (Machado, (2019).
Perceptual Linear Prediction (PLP)	It is a procedure for modeling the human auditory system. It is used to represent the spectral envelope of a speech signal and is often used in speaker recognition, And as a technique for human auditory system modeling (Aljasem, 2021).
Gammatone Filter bank	A technique that models the cochlear filtering of speech signals in the human ear, by applying a set of gammatone filters to the speech signal and then extracting statistical features from the resulting filter outputs (Wang H. &., 2020)
Wavelet Transform	A technique that decomposes the speech signal into a set of wavelet coefficients that capture both spectral and temporal information and then extracts statistical features from these coefficients, (Chunrong, 2007)
Formants Wavelet Entropy (FEW) (Al-Ali, 2021)	A technique that combines the analysis of formants and wavelet entropy. Formants are frequency bands that are most intense in human speech, and they can be used to characterize the unique resonances of an individual's vocal tract.
i-vector and Probabilistic Linear Discriminant Analysis (PLDA) (Al-Ali, 2021)	The i-vector and Probabilistic Linear Discriminant Analysis (PLDA) system is a commonly used technique for extracting a low-dimensional representation of a speaker's voice (known as an i-vector) and then using a PLDA model to classify the speaker as either genuine or impostor. Voice feature extraction-i-vector extraction-PLDA modeling -Speaker verification- Voice analysis (Morrison G. S., 2019).
Relative Spectral (RASTA) Filtering (Franco-Pedroso, 2016)	A technique that models the dynamics of human auditory processing to filter a speech signal. The resulting spectral coefficients are used as features.

Feature Matching Techniques

Table 3:- Techniques utilized for matching of extracted voice features.

Feature Matching Techniques	Description
Gaussian Mixture Models (GMMs) (Haris, 2015)	GMMs are a probabilistic model that represents the distribution of speech features for each speaker as a mixture of Gaussian distributions (Xian Y. M., 2013) .During training, GMMs are fit to the speaker's feature vectors, and during testing, the likelihood of the feature vectors given each GMM is computed. The speaker with the highest likelihood is selected as the match.
Vector Quantization (VQ) (Kaur, 2015)	VQ is a clustering algorithm that partitions the feature space into a set of code vectors. Each feature vector is then assigned to the nearest code vector (Segundo E. S., 2019). During training, code vectors are generated for each speaker, and during testing, the feature vector is assigned to the speaker with the nearest code vector (Segundo E. S., 2017).
Support Vector Machines (SVMs) (Haris, 2015)	SVMs are a machine learning algorithm that learns a boundary between the positive and negative examples in the feature space. During training, SVMs learn a hyperplane that separates the feature vectors of each speaker (Ahmad, 2018). During testing, the feature vector is assigned to the speaker whose hyperplane it lies on.
Deep Neural Networks (DNNs) (Wang Q. Z., Towards Query-Efficient Adversarial Attacks Against Automatic Speech Recognition Systems, 2021; Kim, 2021)	DNNs are a type of artificial neural network that can learn complex nonlinear mappings between the input feature vectors and the speaker labels (Choi, 2020). During training, DNNs learn a set of weights that transform the feature vectors into a speaker representation (Ahmad, 2018). During testing, the speaker with the highest activation for the given feature vector is selected as the match
Hidden Markov Models (HMMs) (Singh, 2017)	Hidden Markov Models (HMMs) are a commonly used feature-matching technique in speaker recognition systems. HMMs are statistical models that can be used to model the probability distribution of a sequence of feature vectors. HMMs are especially useful when the feature vectors are correlated over time, as is often the case with speech signals.
Electrical network frequency (ENF) (Esquef, 2014)	The electrical network frequency (ENF) is the frequency of the alternating current (AC) power grid in a particular region. It is typically 50 Hz or 60 Hz, depending on the country and the power grid. ENF is generated by the power stations and is then distributed through power lines to homes and businesses.

Equal error rate (EER) (Esquef, 2014)	The equal error rate (EER) is a common metric used to evaluate the performance of biometric systems, such as speaker identification systems. The EER is the point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal. The FAR is the proportion of times that the system incorrectly identifies an unauthorized user as an authorized user, while the FRR is the proportion of times that the system incorrectly rejects an authorized user.

Discussion

Forensic voice and speaker recognition are complex fields that require specialized training and expertise (Schilling, 2015). The accuracy of these techniques can be affected by factors such as recording quality, background noise, and changes in the speaker's voice over time. As such, it is important to use these techniques in conjunction with other forensic evidence and investigative methods to build a strong case. The Acoustic analysis involves the examination of audio recordings to extract acoustic features such as pitch, frequency, amplitude, and duration (Batalla, 2014). These features can be used to identify patterns that are unique to a particular speaker and to compare different speech samples to determine whether they were produced by the same person Phonetic analysis involves the examination of the speech sounds and pronunciation patterns used by a speaker (Rao, 2021). This can include the study of phonemes, which are the smallest units of sound in a language, as well as larger units such as syllables, words, and phrases. Phonetic analysis can be used to identify regional or foreign accents, speech disorders, or other unique vocal characteristics that can help to identify a speaker.

In the previous studies the LR (Likelihood Ratio) and various regression models/equations have been utilized for the voice analysis and recognition. The various features for language and voice assessments are taken into consideration such as pitch level, its variability and pattern, voice quality in general, vocal fry and other, variations in intensity, dialect whether its regional, foreign or idiolect, Articulation of vowels, consonants, Misarticulation and nasality, prosody rate, speech bursts etc. along with non-fluencies and speech disorders were assessed (Hollien H. B., 2014). Speaker recognition/identification seems challenging under high level of reverberations and noise and ICA-EBM is used for the separation of speech from noisy voice signals. The i-vector along with Gaussian probabilistic linear discriminant analysis was used as a classifier. Studies related to voice impersonation reveal that these individuals can exploit the differences in mean F0, speech rate, vocal fold pattern and vowel formant distributions to create different voice identities, so for speaker recognition usually a front-end voice disguise system used based on the disguise metric. And the distinction between normal voices and pathological voices by AVQI (Acoustic Voice Quality Index). Various supervector regression methods have also been utilized for forensic voice comparison and it has been concluded that super vector regression techniques have greater validity and reliability than GMM-UBM and GMM-SVM. For text independent speaker recognition, the PNCC and RASTA PLP features are utilized and PNCC results in accurate identification of speaker even for noisy signals and GMM provides better results than I vector method for short utterances (2-3 sec) and for longer utterances (6-9 sec) i vector provides significant results. Vowel quality and quantity studies reveals that the vowel quality of each individual's dialect is distinguishable when compared with the standard as it is unique to the dialectal accent, and it has been proved that considering all the features related to the vowel quality of dialect is more effective for profiling of speakers (Ahuja P. &., 2018). Regional dialectal study in India have shown that with variation in region the vowel quality and quantity along with prescription model of prosody changes, for e.g., long vowel /ʌ/ is used by Bhojpuri, Chhattisgarhi, Kanauji, Marwari, and Haryanvi as compared to Khariboli. For speaker profiling the acoustic features related to sentence intonation and lexicon are unique to the speaker belonging to a regional dialectal group as the changes in dialectal accent is more pronounced in male than the female speakers of a regional dialect (Kulshreshtha, 2012). One of the main challenges in the forensic acoustic and phonetic analysis is the variability of speech. Speech patterns can be influenced by a variety of factors, including age, gender, height, health, emotional state, and even the context in which the speech is produced (Albuquerque L. O.-C., 2020; Cerrato, 2000; Hansen, 2015). Evaluation of deteriorated acoustic circumstances along with recognizing the voice segments/fragments turns out to be challenging. Time interval measure, pitched spectrum analysis, zero-crossing rate, higher statistics of the LPC domain, and amalgamation of various attributes are the factors in voice detection and analysis (Rao, 2021). As a result, forensic linguists must take these factors into account when analyzing speech samples and must use multiple techniques to ensure the accuracy of their findings. Despite these challenges, forensic acoustic and phonetic analysis can be a powerful tool in criminal

investigations, providing valuable evidence that can help to link suspects to crimes and to identify individuals who may have been involved in criminal activity.

Conclusion

Forensic voice analysis and recognition constitutes of analytical approaches such as acoustic-phonetic, acoustic-phonetic-statistical, auditory spectro-graphic and human-supervised-automatic as a commanding/high powered tool for law enforcement and legal executives professionals to identify and analyze the audio recordings/ voice/ speech in criminal and civil investigations. Forensic acoustics has emerged significantly in recent years with advancement in technology and techniques like spectrographic analysis, voice comparison, phonetic analysis, speaker identi-fication and forensic transcription utilizing i-vector and PLDA, PLP (Perceptual Linear Predictive), RASTA, MFCC, HMM, DNN, GMM's that allow more accurate, valid and reliable analysis of audio evidences that link suspects to crimes and to identify individuals who may have been involved in criminal activity. Various studies have been conducted for regional dialects in India that uses Prescription Model of Prosody, vowel quality and quantity for data collection further research can be done for other dialectal accents, that are essential for forensic laboratories dealing with speaker profiling and identification.

References:

1. Ahmad, J. S. (2018). Determining speaker attributes from stress-affected speech in emergency situations with hybrid SVM-DNN architecture. Multimedia Tools and Applications. doi: https://doi.org/10.1007/s11042-016-4041-7
2. Ahuja, P. &. (2018). Forensic speaker profiling: the study of supra-segmental features of Gujarati dialects for text – independent speaker identification. Australian Journal of Forensic Sciences. doi:https://doi.org/10.1080/00450618.2016.1237547
3. Ahuja, P. &. (2018). Forensic speaker profiling: the study of supra-segmental features of Gujarati dialects for text – independent speaker identification. Australian Journal of Forensic Sciences. doi: https://doi.org/10.1080/00450618.2016.1237547
4. Al-Ali, A. K. (2021). Enhanced forensic speaker verification performance using the ICA-EBM algorithm under noisy and reverberant environments. Evolutionary Intelligence,, 1475–1494. doi:https://doi.org/10.1007/s12065-020-00406-8
5. Albuquerque, L. O.-C. (2020). A Comprehensive Analysis of Age and Gender Effects in European Portuguese Oral Vowels. Journal of Voice, 143.e13-143.e29. doi: https://doi.org/10.1016/j.jvoice.2020.10.021
6. Albuquerque, L. O.-C. (2020). A Comprehensive Analysis of Age and Gender Effects in European Portuguese Oral Vowels. Journal of Voice, 143.e13-143.e29. doi:https://doi.org/10.1016/j.jvoice.2020.10.021
7. Aljasem, M. I. (2021). Secure Automatic Speaker Verification (SASV) System Through sm-ALTP Features and Asymmetric Bagging. IEEE Transactions on Information Forensics and Security,, 3524–3537. doi: https://doi.org/10.1109/tifs.2021.3082303
8. Amin, T. B. (2014). Glottal and Vocal Tract Characteristics of Voice Impersonators. IEEE Transactions on Multimedia, 668–678. doi:https://doi.org/10.1109/tmm.2014.2300071
9. Arabi, A. T. (2023). Correlation Between Auditory-perceptual Parameters and Acoustic Characteristics of Voice in Theater Actors. Middle East Journal of Rehabilitation and Health. doi: https://doi.org/10.5812/mejrh-131241
10. Batalla, F. N. (2014). Acoustic Voice Analysis Using the Praat programme: Comparative Study With the Dr. Speech Programme. Acta Otorrinolaringologica (English Edition),, 170–176. doi: https://doi.org/10.1016/j.otoeng.2014.05.007
11. Brockmann-Bauser, M. &. (2023). Do We Get What We Need from Clinical Acoustic Voice Measurements? Applied Sciences,, 941. doi:https://doi.org/10.3390/app13020941
12. Brown, G. &. (2017). Automatic sociophonetics: Exploring corpora with a forensic accent recognition system. Journal of the Acoustical Society of America,, 422–433. doi:https://doi.org/10.1121/1.4991330
13. Cenceschi, S. M. (2021). The Variability of Vowels' Formants in Forensic Speech. IEEE Instrumentation & Measurement Magazine, 38–41. doi:https://doi.org/10.1109/mim.2021.9345600
14. Cerrato, L. F. (2000). Subjective age estimation of telephonic voices. Speech Communication,, 107–112. doi:https://doi.org/10.1016/s0167-6393(99)00071-0
15. Choi, D. K. (2020). Selective Audio Adversarial Example in Evasion Attack on Speech Recognition System. IEEE Transactions on

Information Forensics and Security, 526–538. doi:https://doi.org/10.1109/tifs.2019.2925452
16. Chunrong, X. J. (2007). A Dynamic Feature Extraction Based on Wavelet Transforms for Speaker Recognition. International Conference on Electronic Measurement and Instruments. doi:https://doi.org/10.1109/icemi.2007.4350520
17. De Lara, J. a. (2018). A method to compensate the influence of speech codec in speaker recognition. International Journal of Speech Technology,, 975–985. doi:https://doi.org/10.1007/s10772-018-9547-0
18. Denk, F. D. (2014). Enhanced forensic multiple speaker recognition in the presence of coloured noise. International Conference on Signal Processing and Communication Systems. doi:https://doi.org/10.1109/icspcs.2014.7021056
19. Enzinger, E. &. (2017). Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case. Forensic Science International,, 30–40. doi:https://doi.org/10.1016/j.forsciint.2017.05.007
20. Eskidere, Ö. (2016). Source Digital Voice Recorder Identification by Wavelet Analysis. International Journal on Artificial Intelligence Tools. doi:https://doi.org/10.1142/s0218213016500160
21. Esquef, P. a. (2014). Edit Detection in Speech Recordings via Instantaneous Electric Network Frequency Variations. IEEE Transactions on Information Forensics and Security, 2314–2326. doi:https://doi.org/10.1109/tifs.2014.2363524
22. Franco-Pedroso, J. &.-R. (2016). Linguistically-constrained formant-based i-vectors for automatic speaker recognition. Speech Communication, 61–81. doi:https://doi.org/10.1016/j.specom.2015.11.002
23. Gamer, M. R. (2006). Psychophysiological and vocal measures in the detection of guilty knowledge. International Journal of Psychophysiology, 76–87. doi:https://doi.org/10.1016/j.ijpsycho.2005.05.006
24. Gerlach, L. M. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. Speech Communication, 85–95. doi:https://doi.org/10.1016/j.specom.2020.08.003
25. Goyal, A. S. (2019). Identification of source mobile hand sets using audio latency feature. Forensic Science International, 332–335. doi:https://doi.org/10.1016/j.forsciint.2019.02.031
26. Grillo, E. U. ((2020). An Assessment of Different Praat Versions for Acoustic Measures Analyzed Automatically by VoiceEvalU8 and Manually by Two Raters. Journal of Voice, 17–25. doi:https://doi.org/10.1016/j.jvoice.2020.12.003
27. Hansen, J. H. (2015). Speaker height estimation from speech: Fusing spectral. Journal of the Acoustical Society of America,, 1052–1067. doi:https://doi.org/10.1121/1.4927554
28. Haris, B. C. (2015). Robust Speaker Verification With Joint Sparse Coding Over Learned Dictionaries. IEEE Transactions on Information Forensics and Security,, 2143–2157. doi:https://doi.org/10.1109/tifs.2015.2450674
29. Harnsberger, J. D. (2009). Stress and Deception in Speech: Evaluating Layered Voice Analysis. Stress and Deception in Speech: Evaluating Layered Voice Analysis. Journal of Forensic Sciences,, 642–650. doi:https://doi.org/10.1111/j.1556-4029.2009.01026.x
30. Hollien, H. B. (2014). Issues in Forensic Voice. Journal of Voice, 170–184. doi:https://doi.org/10.1016/j.jvoice.2013.06.011
31. Hollien, H. B. (2014). Issues in Forensic Voice. Journal of Voice,, 170–184. doi:https://doi.org/10.1016/j.jvoice.2013.06.011
32. Hughes, V. &. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. Speech Communication, 218–230. doi:https://doi.org/10.1016/j.specom.2014.10.006
33. Hughes, V. &. (2020). Sharing innovative methods, data and knowledge across sociophonetics and forensic speech science. Linguistics Vanguard,. doi:https://doi.org/10.1515/lingvan-2018-0062
34. Jessen, M. E. (2008). Forensic Phonetics. Langu- age and Linguistics Compass, 671–711. doi:https://doi.org/10.1111/j.1749-818x.2008.00066.x
35. Kaur, K. &. (2015). Performance analysis of text-dependent speaker recognition system based on template model based classifiers. International Conference on Signal Processing. doi:https://doi.org/10.1109/ispcc.2015.7374994

36. Khelif, K. M. (2017). Towards a Breakthrough Speaker Identification Approach for Law Enforcement Agencies: SIIP. European Intelligence and Security Informatics Conference. doi:https://doi.org/10.1109/eisic.2017.14
37. Kim, A. Y. (2021). Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach. Journal of Medical Internet Research, e34474. doi:https://doi.org/10.2196/34474
38. Kulshreshtha, M. S. (2012). Speaker Profiling: The Study of Acoustic Characteristics Based on Phonetic Features of Hindi Dialects for Forensic Speaker Identification. Springer EBooks,, 71–100. doi:https://doi.org/10.1007/978-1-4614-0263-3_4
39. Machado, T. D. ((2019). Forensic Speaker Verification Using Ordinary Least Squares. Sensors, 4385. doi:https://doi.org/10.3390/s19204385
40. Madill, C. &. (2020). Impact of Instructed Laryngeal Manipulation on Acoustic Measures of Voice–Preliminary Results. Journal of Voice, 143.e1-143.e11. doi:https://doi.org/10.1016/j.jvoice.2020.11.004
41. Morrison, G. S. ((2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). Speech Communication,, 242–256. doi:https://doi.org/10.1016/j.specom.2010.09.005
42. Morrison, G. S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. Science & Justice,, 245–256. doi:https://doi.org/10.1016/j.scijus.2013.07.004
43. Morrison, G. S. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. Forensic Science International, 92–100. doi:https://doi.org/10.1016/j.forsciint.2016.03.044
44. Morrison, G. S. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction. Speech Communication,, 119–126. doi:https://doi.org/10.1016/j.specom.2016.07.006
45. Morrison, G. S. (2019). A statistical procedure to adjust for time-interval mismatch in forensic voice comparison. Speech Communication,, 15–21. doi:https://doi.org/10.1016/j.specom.2019.07.001
46. Morrison, G. S. (2023). Forensic Voice Comparison: Overview. Elsevier EBooks,, 737–750. doi:https://doi.org/10.1016/b978-0-12-823677-2.00130-6
47. Nayana, P. K. (2017). Performance comparison of speaker recognition systems using GMM and i-Vector methods with PNCC and RASTA PLP features. International Conference on Intelligent Computing. doi:https://doi.org/10.1109/icicict1.2017.8342603https://doi.org/10.1109/icicict1.2017.8342603
48. Rao, S. (2021). Forensic Aspects of Voice Analysis in India. International Journal for Research in Applied Science and Engineering Technology, 1990-1993. doi:https://doi.org/10.22214/ijraset.2021.37702
49. Schilling, N. &. (2015). Unmasking Identity: Speaker Profiling for Forensic Linguistic Purposes. Annual Review of Applied Linguistics,, 195–214. doi:https://doi.org/10.1017/s0267190514000282
50. Segundo, E. S. (2017). . A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. Journal of Voice,, 644.e11-644.e27. doi:https://doi.org/10.1016/j.jvoice.2017.01.005
51. Segundo, E. S. (2019). The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals. Journal of the International Phonetic Association,, 353–380. doi:https://doi.org/10.1017/s0025100318000130
52. Singh, R. J. (2017). Voice disguise by mimicry: deriving statistical articulometric evidence to evaluate claimed impersonation. ET Biometrics, 282–289. doi:https://doi.org/10.1049/iet-bmt.2016.0126
53. Sondhi, S. K. (2015).). Acoustic analysis of speech under stress. International Journal of Bioinformatics Research and Applications,, 417. doi:https://doi.org/10.1504/ijbra.2015.071942
54. Todkar, S. P. (2018). Speaker Recognition Techniques: A Review. International Conference for Convergence for Technology. doi:https://doi.org/10.1109/i2ct.2018.8529519

55. Univaso, P. (2017). Forensic Speaker Identification: a tutorialForensic Speaker Identification: a tutorial. IEEE Latin America Transactions,, 1754–1770. doi:https://doi.org/10.1109/tla.2017.8015083
56. Walsh, J. K. (2007). Joint Iterative Multi-Speaker Identification and Source Separation using Expectation Propagation. Workshop on Applications of Signal Processing to Audio and Acoustics. doi:https://doi.org/10.1109/aspaa.2007.4393034
57. Wang, H. &. (2020). The application of Gammatone frequency cepstral coefficients for forensic voice comparison under noisy conditions. Australian Journal of Forensic Sciences,, 553–568. doi: https://doi.org/10.1080/00450618.2019.1584830
58. Wang, Q. Z. (2021). Towards Query-Efficient Adversarial Attacks Against Automatic Speech Recognition Systems. IEEE Transactions on Information Forensics and Security, 896–908. doi: https://doi.org/10.1109/tifs.2020.3026543
59. Wang, Q. Z. (2021b). Towards Query-Efficient Adversarial Attacks Against Automatic Speech Recognition Systems. IEEE Transactions on Information Forensics and Security,, 896–908. doi:https://doi.org/10.1109/tifs.2020.3026543
60. Xian, Y. M. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices. Speech Communication, 796–813. doi:https://doi.org/10.1016/j.specom.2013.01.01
61. Xian, Y. M. (2013). Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. Journal of the Acoustical Society of America, EL54–EL60. doi:https://doi.org/10.1121/1.4773223
62. Xian, Y. M. (2013i). Reliability of human-supervised formant-trajectory measurement for forensic voice comparison. Journal of the Acoustical Society of America,, EL54–EL60. doi:https://doi.org/10.1121/1.4773223
63. Xian, Y. M. (2016). Use of relevant data, quantitative measurements, and statistical models to calculate a likelihood ratio for a Chinese forensic voice comparison case involving two sisters. Forensic Science International,, 115–124. doi:https://doi.org/10.1016/j.forsciint.2016.08.017