Neurocomputing 640 (2025) 130334

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Survey Paper

A comprehensive study of audio profiling: Methods, applications, challenges, and future directions

Anil Pudasaini [a,*], Muna Al-Hawawreh [a], Mohamed Reda Bouadjenek [a], Hakim Hacid [b], Sunil Aryal [a]

[a] School of Information Technology, Deakin University, Geelong, Australia
[b] Technology Innovation Institute, United Arab Emirates

ARTICLE INFO

Communicated by J. Andreu-Perez

Keywords:
Audio profiling
User profiling
Speaker profiling
Voice privacy
Privacy preservation
Personality detection
Age detection
Emotion detection
Acoustic event detection
Personality traits detection
Voice pathology detection
Mental health inference
Acoustic scene classification
Gender detection

ABSTRACT

Audio profiling is at the forefront of a technological breakthrough, offering rich insights into human behavior, emotions, physical attributes, and environmental contexts through detailed analysis of voice data. As we embrace an era where the integration of smart technologies equipped with the ability to capture sound is becoming ubiquitous, the capacity to accurately infer personal traits such as age, gender, height, weight, emotional state, personality, and even environmental contexts through voice analysis opens up vast opportunities across law enforcement, healthcare, social and commercial services, and entertainment. This emerging field promises to enhance our interaction with technology by not only understanding who we are but also by interpreting the world around us. However, the remarkable landscape is fraught with challenges, including data imbalances, the complexity of predictive models, and significant privacy concerns regarding the handling of sensitive paralinguistic information. This survey explores deep into the current landscape of audio profiling, examining the techniques and datasets in use, and showcasing its diverse applications while highlighting the need for advanced methodologies, enriched dataset development, and robust privacy preservation techniques.

1. Introduction

Speech serves as a primary mode of human communication, allowing the expression of thoughts, ideas, information, and emotions. Beyond linguistic contents, a speech signal can convey a wealth of information, such as the speaker's gender, age group (e.g., senior, youth, child), and their emotional and physical state (e.g., tired, stressed, intoxicated)–details readily discerned by the human ear. Through an innate skill known as speaker profiling, individuals can instantly recognize these aspects and adapt their responses accordingly, enriching the communication experience. While humans possess this skill naturally, machines have yet to reach the same level of proficiency.

The landscape of voice technology has been transformed with the advent of 'Audio Profiling', a novel concept that marks a significant leap in the field of voice analysis. **Audio profiling** (AP) refers to the process of using voice recordings—acquired from sources such as smart assistants (e.g., Apple Siri, Google Assistant, Amazon Alexa), voice services (e.g., Google search, ChatGPT voice, IBM watsonx), standard phone calls and other recording devices—to infer a range of paralinguistic attributes, circumstances, and environmental contexts, which are then used to profile the speaker's identity and context. The profiling process aims to extract meaningful insights from human voice signals, environmental sounds, and transform them into quantifiable and interpretable representations to understand various aspects of the speaker's background and identity. This practice involves both manual analysis and computational techniques, particularly machine learning and artificial intelligence, to automate the deduction of these parameters from voice data. While terms like "Speaker Profiling" and "User Profiling" are frequently used, "Audio Profiling" encompasses a broader scope. Unlike speaker profiling, which focuses solely on individual speaker characteristics, audio profiling includes environmental awareness, capturing a wider spectrum of auditory information. Profiling methods include perception-based, parameter-based, and computation-based approaches [1]. For the scope of this survey, we focus on **Computational**

* Corresponding author.
E-mail addresses: s223786275@deakin.edu.au (A. Pudasaini), muna.alhawawreh@deakin.edu.au (M. Al-Hawawreh), reda.bouadjenek@deakin.edu.au (M.R. Bouadjenek), hakim.hacid@tii.ae (H. Hacid), sunil.aryal@deakin.edu.au (S. Aryal).

https://doi.org/10.1016/j.neucom.2025.130334
Received 3 July 2024; Received in revised form 28 March 2025; Accepted 21 April 2025
Available online 2 May 2025

Profiling for the scope of this survey. This approach uses computational techniques, particularly machine learning, and deep learning, to analyze voice signals and extract patterns or features associated with specific characteristics.

The introduction of audio profiling as a term and a concept is crucial in the current technological landscape. With voice recording technologies becoming ubiquitous in our daily lives, understanding the nuances and implications of audio profiling is paramount. Voice data, inherently rich in information, extends beyond simple linguistic content to include a multitude of paralinguistic features. Features such as stress, perturbation, timbre, tempo, pitch, rhythm, and intonation range, offer valuable insights into an individual's emotional state, age, gender, mental health, physical health, personality traits and even socio-economic status [2]. The ability to infer such sensitive attributes from voice analytics has immense potential in applications ranging from healthcare to personalized services. The domain of Audio Profiling is broad. Several studies study the inference of individual attributes by leveraging features extracted from voice recordings [3–5]. However, few studies provide a holistic view of the field. Moreover, to the best of our knowledge, privacy concerns related to audio profiling have not been adequately addressed. Consequently, there exists a substantial research gap regarding the multiple paralinguistic features that can be inferred from voice and the associated privacy implications. Addressing this gap is essential, especially as voice interactions increase exponentially with the rise of new digital technologies, making it imperative to understand and resolve privacy concerns.

This literature review aims to dissect the multifaceted nature of audio profiling. It intends to scrutinize the methods and effectiveness of extracting personal attributes from voice data and explore potential use cases. Additionally, the survey presents a comprehensive overview of the state-of-the-art machine and deep learning techniques for inferring sensitive attributes from voice recordings and performing audio profiling. Furthermore, it aims to shed light on potential key issues associated with audio profiling and explore viable solutions. The following are the major contributions of this survey article:

1. **Comprehensive Literature Review**: We present a broad review of literature relating to attributes—age, gender, emotion, mental health, personality traits, voice pathology, acoustic scene classification, acoustic event detection under a single canopy.
2. **Dataset Comparison**: A collection of datasets for each attribute inference task and their comparison along different aspects.
3. **Use Cases**: We provide a dedicated section highlighting the immense potential applications of Audio Profiling across various sectors such as commercial, health, law enforcement, forensics, and entertainment.
4. **Future Directions**: We outline potential research challenges, shedding light on promising avenues for exploration, and discuss the implementation of privacy measures aimed at protecting sensitive attribute inference in voice analytics.

This literature review covers specific attributes, the challenges impeding research, unaddressed privacy concerns, and discusses potential avenues for workable solutions to address these concerns. It also highlights the potential of audio profiling with major use cases within the current technological landscape.

We initiated the selection of articles by conducting searches for recent research articles using a limited number of keyword seeds on prominent platforms such as Scopus, Arxiv, IEEE Xplore, Google Scholar, ACM Digital Library, Web of Science, and others platforms. On top of that, we also reviewed articles collected in the social and behavioral sciences from sources such as APA (American Psychological Association) journal. The search queries were formulated by combining keywords related to Speaker/User profiling, attributes of interests like age, gender, height, weight, emotions, voice pathology, environmental aspects, personality detection from voice, paralinguistic features extraction and privacy preservation.

To ensure comprehensive coverage, we expanded the initial set of articles by including those cited by or citing articles within this set. We further augmented our selection by incorporating noteworthy research featured in books and media to encompass a broad range of significant publications in the field. This iterative process continued until no new articles were discovered. We then engaged in discussions to analyze the relevance and importance of the selected articles by reviewing abstracts and main findings. Papers deviating from the survey's scope were excluded during this evaluation. The chosen papers constitute the core of this survey. To maintain currency, we continuously updated our selection throughout the survey writing process to include recently published works.

The rest of this paper is organized as follows: First, Section 2 compares our work with other related works. Following this, Section 3 discusses background information on Audio Profiling. Next, in Section 4, general audio profiling pipeline is discussed. Following it, Section 5 discusses in-depth literature, methods and datasets used for individual AP tasks. Following it, Section 6 presents information and comparisons of prominent datasets on Audio Profiling. Next, Section 7 showcases various use cases and applications. Following it, Section 8 discusses critical issues, open research challenges, and future directions. Finally, Section 9 concludes the survey.

2. Comparison with other surveys

In the realm of voice inference, the landscape of existing research exhibits a scattered mosaic of studies exploring distinct attributes such as gender, age, accent, region of geographical origin, emotion, mental health, voice pathology, personality traits, and acoustic scene classification. Despite the abundance of individual research, a comprehensive study that consolidates these distinct aspects under one unified canopy is noticeably lacking. Our survey endeavors to bridge this gap by providing a holistic overview, integrating various attributes, methodologies, and applications that have not been collectively reviewed to date. This survey serves as a one-stop resource for understanding audio profiling and the various attributes derivable from voice recordings, including feature extraction, learning models, and datasets. In addition, our survey incorporates a dedicated section on privacy preservation.

Due to the absence of surveys similar to ours, a one-to-one comparison is not possible. However, we have selected specific elements as the basis for comparison with other surveys/review articles. Table 1 showcases the comparison of our survey with other studies based on the selected criteria.

- **Attributes Covered**: Refers to the number of attributes discussed in the literature. Attributes can either be personal or environmental.
- **Feature Extraction**: Examines the various signal domains and methods used for feature extraction, shedding light on the techniques employed in audio profiling.
- **Use Cases**: Involves the detailed exploration of potential use cases.
- **Datasets**: Refers to the comparison of datasets available, based on different aspects of the dataset.
- **Learning Models**: Covers the use of learning models, which may include statistical models, machine learning models, or deep learning models. These models can either be fully covered (all three types are used) or partially covered (only one or two types are used).

Differences from Existing Surveys and Reviews: Our comprehensive survey spans a wide array of attributes derivable from voice, including age, gender, emotion, mental health, voice pathology, and more, offering a broad overview that integrates multiple disciplines and methodologies. While focused surveys and review papers provide valuable insights and methodologies tailored to specific attributes like

Table 1
Comparison with other studies.

Ref.	Attributes covered	Year	Feature extraction		Dataset comparison	Learning models	Use cases
			Domain	Methods			
[6]	Mental health	2018	○	○	◑	●	○
[7]	Emotion	2019	○	○	◑	●	◑
[8]	Voice pathology	2019	◑	◑	◑	◑	○
[9]	Event detection	2019	○	○	●	●	◑
[10]	Scene classification	2020	○	◑	○	◑	◑
[11]	Emotion	2021	◑	●	◑	●	○
[12]	Personality, Gender	2021	○	○	○	◑	◑
[13]	Voice pathology	2022	◑	◑	◑	●	◑
[14]	Age, Gender, Height	2023	●	◑	○	◑	◑
[15]	Emotion	2023	●	◑	●	●	◑
Our work	All listed attributes	2024	●	●	●	●	●

Notes: ●– *Fully Considered*; ◑– *Partially Considered*; ○– *Not Considered*.

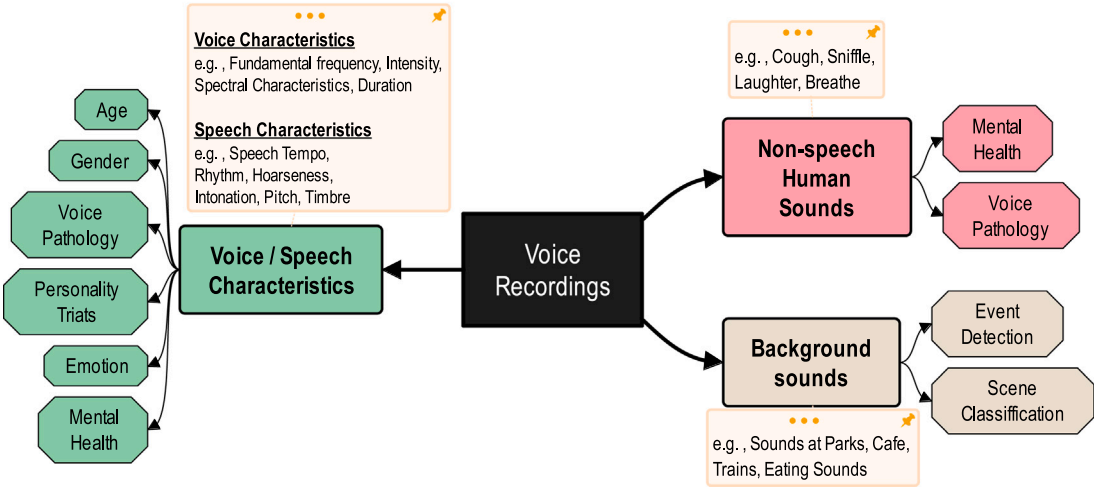


Fig. 1. Overview of attributes evident in voice data.

emotion and voice pathology, our survey connects these diverse strands of research. By adopting a broader perspective, we highlight how different areas of study intersect and influence one another. These connections are not limited to a single discipline but extend across various fields. For instance, several studies [16,17] in emotion recognition rely on gender recognition, as emotion and gender are inextricably intertwined.

3. Background

3.1. Overview of audio profiling

Audio profiling encompasses signal processing, machine learning, deep learning, and cognitive psychology. The field has evolved from early speech recognition research to advanced computational methods capable of extracting complex patterns from voice recordings. With the advent of machine learning (ML) and deep learning (DL), profiling techniques have become significantly more sophisticated. These audio profiling-based models can automatically infer speaker attributes, such as age, gender, and emotional state, by learning patterns from large-scale voice datasets, significantly enhancing the accuracy and breadth of applications, ranging from improving security and healthcare services to enriching user experiences in entertainment and social media, as discussed in more detail in Section 7.

Audio profiling is an invaluable tool, and its true potential lies in the depth of information that can be extracted from voice recordings. These voice recordings usually contain information about the recorded speakers and their context, encompassing speech, non-verbal human sounds, and environmental background sounds. For instance, the analysis of human voice in these recordings enables the estimation of speaker traits, such as age, gender, and physical attributes [18]. ML and DL [19] models trained on speech data or voice recording have demonstrated impressive accuracy in inferring these characteristics by analyzing pitch, formant frequencies, and vocal tract length, as it will be detailed in Section 4. Beyond these physical traits, voice profiling also extends to psychological and social attributes. Research has shown that voice attractiveness correlates with objective measures of physical appeal, such as waist-to-hip ratio in women and shoulder-to-hip ratio in men [20]. Furthermore, individuals with more melodious or resonant voices are often perceived as having favorable personality traits, influencing social interactions [21]. The role of voice in reproductive viability has also been explored, with studies linking vocal changes during puberty and menopause to cues about sexual maturity and reproductive status [22]. These findings highlight the broad applicability of voice analysis in understanding human behavior and physiology.

In addition to the human voice, background sounds in voice recordings can capture valuable contextual information about the speaker's environment. An active area of research, acoustic scene classification, focuses on categorizing sounds to deduce the context [23]. Specific applications include detecting distinct sounds within recordings, such as heavy machinery operating or running water in the background [9]. Additionally, voice recordings often contain non-speech human sounds,

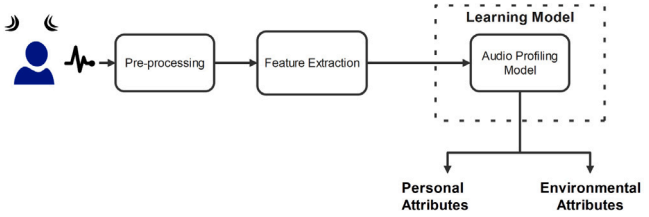


Fig. 2. Audio profiling pipeline.

such as sniffing or coughing, which can be detected and used to further profile the speaker [24]. Fig. 1 provides an overview of the attributes evident in voice recordings.

3.2. Rationale and techniques behind audio profiling

The concept of audio profiling holds immense potential for a wide range of applications. The ability to infer personal characteristics such as age, gender, emotional state, presence of voice pathologies, mental health status, and personality traits, as well as contextual factors like indoor or outdoor environments from audio data, represents a compelling area of research. The far-reaching implications of this technology in our increasingly technology-driven society underscore the need for further exploration and development of audio profiling techniques. Extracting such a diverse array of information solely from the audio modality enables numerous practical applications across various domains, highlighting the importance of continued research in this field.

Audio profiling can be utilized for voice forensics, face reconstruction, or suspect tracking by law enforcement agencies. It can also enhance human–machine interaction by enabling systems to better understand user characteristics and ambient context. Section 7 explores additional use cases, while the inference of personal or environmental attributes is thoroughly examined in Section 5.

4. Audio profiling pipeline

The Audio Profiling (AP) pipeline is generally divided into three parts, as shown in Fig. 2: Data Acquisition and Pre-processing, Feature Extraction, and Learning Model. Our survey is organized around these stages of audio profiling. We begin by discussing the preliminary stages of audio profiling, starting with Data Acquisition and Pre-processing in Section 4.1. Next, we provide a detailed study of feature extraction in Section 4.2. The components of feature extraction are illustrated in Fig. 4, and we also group the features in different domains and the extraction methods employed, as shown in Fig. 4(b). Furthermore, in Section 4.3, we examine various learning models used in audio profiling. Finally, we explore the literature, methods, datasets, and the evolution of research conducted for each AP task in Section 5.

4.1. Data pre-processing

Pre-processing is the first step of audio profiling after collecting the relevant data from available sources. It encompasses various techniques and steps that depend on the dataset and the nature of the targeted task. For instance, audio data collected from different sources may have non-uniform recording settings, with variations in audio channels or sampling frequencies. These variations can be mitigated by converting audio signals into a uniform format through *down-mixing* to a fixed number of channels and re-sampling to a fixed sampling frequency [25]. Additionally, some audio data require noise and interference reduction, achieved through filtering and denoising, to focus on relevant signal aspects. In noisy environments, noise suppression methods can dampen the interference of environmental noise during audio analysis [26], while overlapping sounds can be addressed using sound source separation methods [27].

To further ensure consistency in audio data processing, pre-processing may also include the standardization and normalization of formats across various sample rates and resolutions. Resampling is used to standardize sample rates, enhancing computational efficiency and ensuring compatibility with different models. Normalization adjusts signal amplitudes to prevent bias, while handling variable lengths through segmenting or padding ensures uniform inputs for models that require fixed-length data. These steps are critical in preparing audio data for subsequent analysis, ensuring consistent and optimized inputs for the best performance of profiling algorithms.

Beyond these core techniques, the pre-processing phase can include additional methods, such as applying windowing functions, pre-emphasis, resampling, dBA weighting of the magnitude spectrum, autocorrelation functions, mean–variance normalization, range normalization, delta-regression coefficients, and various vector operations. However, most existing research focuses on three key pre-processing techniques, as illustrated in Fig. 3:

- **Pre-emphasis:** This initial step increases the magnitude of higher frequency sounds relative to lower frequencies. It is typically implemented using the formula:

$$y_t = x_t - \alpha x_{t-1} \tag{1}$$

  where x_t represents the sample signal at time t, x_{t-1} is the previous sample, α is a weight factor, and y_t is the resulting emphasized sample. Pre-emphasis enhances the signal-to-noise ratio by amplifying higher frequencies, which often carry more informative features in speech signals.
- **Framing:** Speech signals vary over time and are non-stationary. To analyze them effectively, they are segmented into short frames, assuming stationarity within each frame. Frames are typically 20–30 ms long, with a 50% overlap between consecutive frames. Within each frame, a spectral feature vector is extracted.
- **Windowing:** Segmenting signals into frames can introduce discontinuities at frame edges, leading to **spectral leakage**. To mitigate this, a window function is applied to each frame. Several windowing functions exist, such as Rectangular, Hamming, Hann (raised cosine), Gauss, Sine, Triangular, Bartlett, Bartlett-Hann, Blackman, Blackman-Harris, and Lanczos. A common choice is the Hamming window, defined as:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad \text{where } 0 \le n \le N-1 \tag{2}$$

  where N is the length of the frame and $n = 0, 1, 2, \ldots, N-1$. The window function tapers the signal at the edges of each frame, reducing spectral leakage.

4.2. Feature extraction

Audio feature extraction refers to the process of capturing and quantifying relevant characteristics or attributes from an audio signal. It involves transforming raw audio waveforms into numerical representations that effectively capture important information for various audio processing tasks. Feature extraction is essential because raw audio waveforms are high-dimensional and contain vast amounts of data. By extracting relevant features, the dimensionality of the audio data is reduced while retaining the essential information required for subsequent analysis or modeling. The primary goal of feature extraction in audio profiling is to represent **variable-size utterances** as fixed-size feature vectors suitable for further processing in model training and result inference. Since audio signals exist in different domains and various features can be extracted based on these domains and using different methods, this section primarily discusses feature extraction research categorized by signal domains and extraction methods (see Fig. 4).

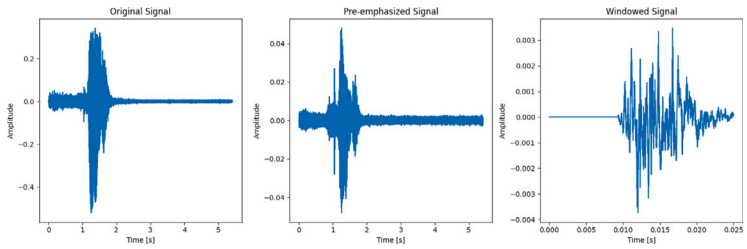


Fig. 3. General pre-processing steps.

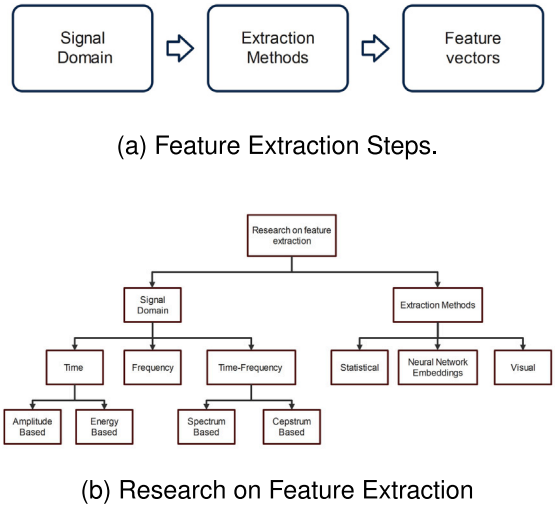


Fig. 4. Feature extraction components.

4.2.1. Signal domains

In this section, we discuss feature extraction from audio signals based on the specified signal domains, including time domain, frequency domain, and time–frequency domain.

Time domain:. The time domain represents audio signals based on amplitude variations over time. Temporal speech characteristics, such as intonation, pause duration, speed, and rhythm, are unique to each individual's speech [28]. By performing waveform analysis on signals in the time domain, distinct patterns related to the speaker's vocal tract and articulation style can be observed. The energy distribution over time, such as variations in loudness and the energy envelope of speech, is a significant identifier. In the time domain, these energy dynamics can be directly observed and measured, offering important clues for speaker identification. For instance, a speaker's level of excitement or stress can often be inferred from these features, as temporal speech characteristics change under such conditions.

Time domain features, such as the Zero-Crossing Rate (ZCR) of a recording, can provide clues about the recording environment. Higher energy levels and varying ZCR values can indicate a noisy or dynamic environment, while lower, steadier values might suggest a calmer setting. Time domain features can be amplitude-based, such as Attack Decay (AD), Attack Decay Sustain Release (ADSR), Log Attack Time (LAT), and Shimmer, or energy-based, such as Root Mean Square (RMS) energy and Short-Time Energy. Other features include rhythm-based metrics, auto-correlation-based features, and ZCR.

Frequency domain:. In the frequency domain, audio signal analysis shifts focus to the spectrum of frequencies present in a sound. The transformation of time-domain signals into the frequency domain is performed using the Fourier Transform. This process enables the examination of the frequency components of both continuous and discrete time-domain signals, simplifying mathematical analysis of the underlying system. The frequency domain provides insights into the harmonic content and timbre of the audio, which are essential for understanding speaker characteristics and environmental contexts.

Some key frequency domain features include Spectral Centroid, Spectral Bandwidth, and Spectral Roll-off. These features are crucial in speaker profiling, as they can reveal unique aspects of a speaker's voice, such as nasality or sharpness. In terms of environmental inference, frequency domain features help identify the type of setting (e.g., indoors vs. outdoors) or the presence of specific background noises (e.g., traffic or machinery) [9].

Time–frequency domain:. The time–frequency domain approach in audio signal processing provides a comprehensive analysis of audio signals by capturing both their temporal and spectral characteristics. This domain is particularly important for complex auditory phenomena where understanding the evolution of frequencies over time is crucial. Time–frequency analysis is achieved through a time–frequency distribution (TFD), resulting in a time–frequency representation (TFR). While the time domain illustrates amplitude changes over time and the frequency domain provides frequency information but lacks temporal details, a TFR bridges this gap by offering both time and frequency resolution. This enables a more detailed understanding of audio signals, allowing for the perception and analysis of the complex interplay between pitch and timing that defines the unique character of sounds. The Short-Time Fourier Transform (STFT) is a widely used method for obtaining a TFR.

Examples of time–frequency features include Band Energy Ratio (BER), Spectrograms, Mel-Spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs), Perceptual Linear Prediction (PLP), Gammatone Cepstral Coefficients (GTCCs), and Chroma Features. Among these, **MFCCs** have been extensively used to identify personal attributes and analyze various environmental sounds. Additional details on these audio features can be found in the study by [29].

4.2.2. Extraction methods

In this subsection, we discuss different feature extraction methods, as illustrated in Fig. 4(b). These include statistical methods, neural network embedding methods, visual methods, and other approaches commonly found in the literature.

Statistical methods:. Statistical methods in feature extraction from audio signals typically involve calculating various statistical measures to summarize the characteristics of the signal. Time domain features such as ZCR, RMS Energy, Temporal Centroid, Attack Time, and Decay Time are calculated using statistical methods over short-time frames. Statistical methods include measures like Mean, Variance, First Quartile (Q1), Third Quartile (Q3), Interquartile Range (IQR), Kurtosis, and Skewness of an audio signal.

Mean reflects the average amplitude, providing insights into the general loudness of the speech or sound environment. **Variance** and **Standard Deviation** indicate the variability or dynamic range of the audio, which are useful for distinguishing between steady and fluctuating sound environments. **Skewness** reveals the asymmetry of the audio

signal's amplitude distribution, which may indicate certain types of speech patterns or environmental sounds. Similarly, **Kurtosis** indicates the presence of outliers or extreme variations in the audio, which might signify unusual speech events or environmental sounds. These features are useful for basic analysis and classification tasks in audio processing, providing insights into the signal's distribution and variability characteristics. In some cases, entire personal attribute inference tasks, such as gender detection, can be performed using statistical measures alone [30].

Low-dimensional representations, such as sequences of features, super-vectors of Gaussian mean components, or low-dimensional vectors like i-vectors (identity vectors) [31], are also extracted using statistical methods like **Joint Factor Analysis**. Applications of these representations include speaker verification, speaker diarization, language recognition, and emotion recognition. A detailed roadmap of how audio features are extracted using statistical methods is provided in [32].

Neural network embeddings:. Neural network embeddings are feature representations learned by neural networks during training. These embeddings are derived from the input data and are optimized to capture the most relevant features for the specific task. Neural networks learn these embeddings as part of their weight parameters, typically in the initial layers of the network. The embedding process usually consists of two parts. First, an encoder network extracts frame-level representations from acoustic features such as MFCCs or filter banks. This is followed by pooling methods, such as global temporal pooling, which aggregate the frame-level representations into a single vector per utterance. This vector can then be used for different training objectives. Various types of neural network embeddings are used in audio processing, including Audio Word Embeddings [33], Speaker Embeddings [34], and Environmental Embeddings [35].

Of particular interest are speaker embeddings, which capture the unique characteristics of a speaker's voice. A prominent example is x-vectors [34], which are extracted using a deep neural network (DNN) architecture that encodes both speaker and phonetic content. X-vectors have been highly effective in tasks such as speaker verification, diarization, and language recognition [36]. Additional details on different encoder architectures and embeddings are discussed by the authors of [37].

Visual methods:. This category of feature extraction emphasizes visual representation methods that convert audio signals into visual formats, such as spectrograms, from which meaningful features can be extracted. **Spectrograms** provide a time–frequency representation, capturing temporal patterns and frequency events. **Mel-spectrograms**, adapted through the Mel scale [38], emphasize perceptually significant frequencies, enhancing tasks like speech and music recognition. **Chromagrams** focus on pitch classes, aiding in music analysis by identifying harmonic structures.

From these visual representations, various features can be extracted, such as mel-spectrogram features and chromagram features. Convolutional Neural Networks (CNNs) and their variants are commonly used to automatically learn and extract relevant features from these visual representations for further analysis and inference. Audio profiling tasks such as gender recognition [39], emotion recognition [40], acoustic scene classification, and event detection have been successfully performed using visual methods.

Apart from these, there are various **end-to-end approaches** that handle all aspects of the audio profiling process, from the initial input to the final output, without requiring manual intervention or separate processing stages. Raw waveforms are used as input, and different audio profiling tasks such as age estimation [41,42], voice pathology detection [43], Speech Emotion Recognition (SER) [44], and acoustic scene classification [45,46] are directly obtained as output. End-to-end systems perform **automatic feature extraction**, learning directly from raw inputs without manual intervention. This approach allows the learning model to uncover and utilize intricate patterns and relationships within the data. Beyond automatic feature extraction, feature sets for audio profiling can also be selected manually by domain experts and researchers based on experiments or using specialized toolkits such as OpenEar, PRAAT, or openSMILE.

4.3. Learning model

This section discusses various learning models for audio profiling. The process of teaching a computer to make decisions or predictions based on data is commonly referred to as "model learning". It involves selecting a model (an algorithm or mathematical representation), training it on a dataset (allowing the model to discover patterns in the data), and evaluating its performance on new, unseen data. Depending on the type of prediction being made, model learning can encompass various tasks beyond classification and regression. In this section, we explore different types of models and architectures based on how they infer personal or environmental attributes. We categorize these models into three groups: Statistical Models, Machine Learning Models, and Deep Learning Models.

4.3.1. Statistical models

Statistical models are grounded in probability theory and statistical inference. They analyze patterns in data by assuming that the data is generated from a specific statistical distribution. For speech recognition tasks, it is necessary to model the distribution of feature vector sequences. Models such as **Gaussian Mixture Models** (GMMs) are commonly used for density estimation and are particularly adept at handling data with multiple sub-populations. This makes them suitable for speaker recognition, where different speech patterns can be modeled as distinct distributions. **Hidden Markov Models** (HMMs) excel in time-series analysis, which is critical for understanding sequential patterns in speech. HMMs have been applied to speaker modeling and can be used to estimate personal attributes such as age [47]. **Universal Background Models** (UBMs) are large GMMs trained on comprehensive datasets to provide independent models for feature distributions. UBMs offer a general model of speech that can be adapted to specific speakers [48]. These models are fundamental for tasks that require capturing the uncertainty and variability inherent in audio data. GMM-based systems have demonstrated their effectiveness in identity verification and related applications. They are used as foundational models in tasks such as age estimation, gender recognition, and language recognition [3,49,50].

4.3.2. Machine learning models

Machine learning models involve algorithms that learn from data to make predictions or decisions. This learning can be supervised (with labeled data) or unsupervised (without labeled data). Supervised Learning Models like **Support Vector Machines (SVMs), K-Nearest Neighbors (KNNs)** are particularly effective in classification tasks. In the world of audio processing, they can be used to distinguish between different speakers (speaker recognition), identify personal attributes such as emotions [51], voice pathology [52] or to identify specific sound scenes and events in an environment (like a fire alarm or bird chirping, or indoor/outdoor environments) [25]. Unsupervised Learning Models like **K-means or hierarchical clustering** are used to group similar audio patterns without prior labeling and segregate different sound sources in the environment. Similarly, there are various machine learning models for regression such as Linear Regression and its variants, Support Vector Regression (SVR) that have been used for audio profiling.

Machine learning models in audio processing are tasked with extracting meaningful information from complex audio signals. They rely on robust feature extraction methods. Key methods include: statistical methods calculating statistical measures to summarize signal characteristics; neural network embeddings learning compact representations via

temporal pooling and attention, such as x-vectors encoding speaker and phonetic information for speaker recognition; and visual representation methods converting audio to spectrograms, mel-spectrograms, and chromagrams, from which features like mel-spectrogram and chromagram features are extracted for tasks like speech, gender, and emotion recognition. Models are then trained to recognize patterns that are indicative of different audio classes or to understand the structure in the data. The choice between supervised and unsupervised methods depends on the nature of the task and the availability of labeled data.

4.3.3. Deep learning models

Deep learning models have significantly advanced the field of machine learning, particularly in handling complex and high-dimensional data like images, audio, and text. Depending upon the task at hand, there are numerous deep learning models that are employed to understand and solve the problem. Deep Learning Models can be supervised and unsupervised, and are adept at handling various tasks, including both classification and regression.

In supervised learning, **Convolutional Neural Networks (CNNs) and their variants** are extensively used for tasks that involve classification and regression, such as identifying emotions, personality or specific environmental settings from audio data. **Recurrent Neural Networks (RNNs)** and their advanced variants like **Long Short-Term Memory networks (LSTMs)** and **Gated Recurrent Units (GRUs)** are particularly effective for sequential data. These models excel in speech recognition tasks, making them ideal for detecting voice pathology, emotional states, and even events in environmental audio.

On the unsupervised side, **Autoencoders** are employed for feature learning and dimensionality reduction. This capability is crucial in understanding intricate voice characteristics or environmental sounds without predefined labels. **Generative Adversarial Networks (GANs)** are known for their generative abilities, they have been used in feature extractions and also to create synthetic audio samples, which can enhance speech signals and contribute to the robustness of training datasets. Some models, like **Transformers and Deep Belief Networks (DBNs)**, are versatile and find applications in both supervised and unsupervised learning contexts. Transformers have gained prominence for their effectiveness in handling sequential data, making them a powerful tool for advanced speech recognition tasks and environmental sound classification. DBNs, with their generative modeling capabilities, are pivotal in complex feature extraction, applicable in scenarios like detecting voice pathology or analyzing personality traits from speech patterns.

All the types of learning models can use either classification/regression or a combination of both as in [41]. Classification and Regression are the two different types of tasks for Audio Profiling. Classification models can also be divided three types. The first type is **Binary Classification** where the classification model assigns labels from a set of pre-existing labels, for instance, in Gender classification, either male or female labels are assigned to the voice recording. Similarly in Mental Health inference, different mental illness are present or absent. The next type is **Multi-Class Classification** where the classification model assigns a single label out of multiple classes, for instance, in Age classification, a sample belongs to one of the several age groups (young, adult, senior) present. Also, in case of Acoustic Scene Classification, a sample belongs one of the different scenes like parks, transport, cafe. The third one is **Multi-Label Classification** where a sample belong to multiple classes at the same time. For instance, in Audio tagging, the identification of presence of specific sounds such as dog barking, car honking, baby crying etc. Similarly, emotional state identification of speaker in audio recording, such as happy, sad, angry, excited etc. On the other hand, regression models predict a numerical value which can be done in case of Age prediction, Height Prediction and Weight estimation. After feature extraction on the voice data, the model is trained on the training dataset and tested on the test set.

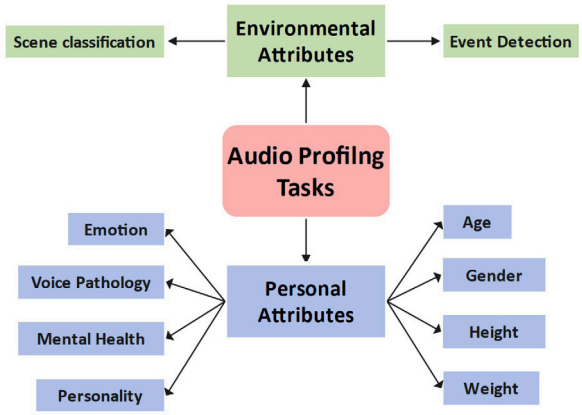


Fig. 5. Audio profiling tasks.

Performance of the audio profiling models are done using the standard metrics : **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** for regression, **Accuracy** and **Unweighted Accuracy (UA)** for classification. In audio profiling, the final outcomes include inferred personal attributes and the deduced environmental context.

5. Audio profiling tasks

Audio profiling tasks involve inferring different personal and environmental attributes. AP integrates audio processing and machine learning to analyze speech recordings beyond conventional speech-to-text applications. It involves extracting multifaceted personal and environmental characteristics such as age, gender, height, weight, emotions, and mental states. The scope of audio profiling extends to environmental context detection as well. The technology can infer whether a recording is captured indoors or outdoors, and can identify specific scenarios such as cafes, public transport, temples, parks, and other outdoor/indoor settings. Similarly, Audio Event Detection (AED) allows the detection of important sounds like doorbells, alarms, or a child who needs attention. Audio profiling tasks are classified based on the personal and environmental attributes they infer, as shown in Fig. 5.

A comprehensive work summing up the methods, algorithms, features on personal attribute inference, and environmental context cues is provided in Sections 5.1 and 5.2.

5.1. Personal attributes inference

The human voice possesses remarkable potential due to its diverse characteristics. From tone and pitch to emotional cues and individual traits, the voice serves as a valuable source of information. This multidimensional nature of speech opens up numerous possibilities across various domains, contributing to the development of advanced technologies and enhancing human–machine interactions. Inferring an individual's age solely from their voice is fascinating; combining this with attributes like gender and emotion provides deeper insights into an individual's behavior and choices. Further augmenting these inferences with the detection of personality traits transforms this information into a treasure trove for digital marketers. Imagine the potential of machines understanding mental states to adapt their interactions accordingly or even diagnosing pathologies from voices. These capabilities underscore the potential of audio profiling to infer personal attributes.

In this section, we discuss a specialized literature review that uncovers deeper insights and key developments in audio profiling. The focus is on analyzing finer details often overlooked in broader surveys, with an emphasis on the methodologies, findings, and contributions of significant studies.

5.1.1. Gender inference from voice

Gender recognition involves analyzing voice signals to determine the gender categories of speakers. The acoustic characteristics of human speech vary by gender due to physiological differences in the glottis and vocal tract. Gender detection is often combined with age group classification, leveraging the fundamental frequency (f0), as male speakers typically exhibit lower frequencies compared to female speakers.

Among the numerous studies on gender detection, we present some of the major ones here. The authors of [53] introduced a two-layer **classifier fusion** technique for automatic gender identification (AGI). The first layer employs acoustic classification with divisive clustering to group speakers based on similar vocal articulatory characteristics. For enhanced performance, the second layer integrates GMM, SVM, and MLP classifiers. The proposed method achieved a 96.53% accuracy on the OGI multilingual corpus [54], surpassing traditional AGI approaches. The authors of [55] proposed gender classification based on **utterance intensity** using Simpson's rule. The area under the normalized curve—obtained by multiplying a factor with a third-degree polynomial fitted through the peaks in each frame of speech utterance (20 ms in length)—was measured to calculate voice intensity. Experiments demonstrated an accuracy of 96.44% on the DARPA TIMIT dataset. To enhance human–computer interaction by integrating speech recognition technology, the authors of [56] extracted Mel coefficients and their derivatives from Hindi vowel speech samples. They achieved an impressive 93.48% accuracy using a combined Support Vector Machine (SVM) and neural network classification approach. The study aimed to identify essential features for gender recognition and assess performance based on the first Mel coefficient. Additionally, the authors of [57] highlighted the importance of other voice signal characteristics, such as pitch and energy, in gender classification. Their method achieved an accuracy of 96.45% using their own datasets.

The study [58] presents a new method for gender recognition using **audio speech features**. It involves data pre-processing for noise reduction, followed by a multi-layer architecture for feature extraction. Fundamental frequency, spectral entropy, spectral flatness, and mode frequency were computed in the first layer, while the second layer employed linear interpolation and Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction after Z-score normalization. SVM and K-Nearest Neighbors (KNN) were used for classification. The model achieved a peak accuracy of 96.8% with K-Nearest Neighbors (KNN) on the TIMIT dataset. The authors extended their work by adding a third layer that calculates Linear Predictive Coding (LPC) coefficients. These three layers were combined for training. The study also introduced a combined dataset for detecting both gender and the geographical region of the speaker, which includes multiple datasets: TIMIT, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and BGC, a self-constructed dataset in random order. TIMIT was primarily used for gender recognition, while RAVDESS was used for emotion detection. The proposed multi-output-based 1D CNN model achieved an accuracy of 93.01% for gender detection, which is comparable to other state-of-the-art models. Additionally, it achieved an accuracy of 97.07% for region detection. However, the study's limitation lies in the relatively small dataset used for region detection [59].

The study discussed in [60] employs a multilayer perceptron (MLP) deep learning model, leveraging acoustic characteristics of voices and speech to ascertain gender. A total of 22 acoustic parameters were measured on acoustic signals, achieving a significant accuracy level of 96.74%. Similarly, [61] conducted a study aiming to predict gender from speech by employing Gaussian Mixture Models (GMM), Multilayer Perceptron (MLP), Vector Quantization (VQ), and Learning Vector Quantization (LVQ). Their predictive model achieved an accuracy of 96.4% using the IViE corpus dataset [62]. Likewise, [63] conducted research to predict gender from speech using **Weighted Supervised Non-negative Matrix Factorization** (WSNMF) and age using a **Generalized Regression Neural Network** (GRNN). Their predictive model achieved a gender recognition accuracy of 96% with the Dutch database [64]. The study in [65] employs deeper Long Short-Term Memory (LSTM) networks to predict gender from audio data, achieving a 98.4% accuracy on the Voice Gender dataset. This study compares its success rate to conventional machine learning methods, highlighting its pioneering role in effective and fast gender detection. In another effort, the authors of [66] focus on increasing the accuracy of machine learning models for gender recognition by utilizing a new ensemble semi-supervised self-labeled algorithm called iCST-Voting. This algorithm integrates an ensemble of popular self-labeled algorithms—Self-training, Co-training, and Tri-training—using them as base learners along with an ensemble of classifiers. Their approach achieved an accuracy of 98.42%, surpassing state-of-the-art supervised algorithms trained with 100% of the training set on the Voice Gender dataset.

The study by [67] strongly supports Deep Neural Networks (DNNs) as an effective choice for gender detection, presenting compelling evidence. By comparing various DNN architectures such as Convolutional Neural Networks (CNNs), Temporal Convolutional Networks (TCNs), Convolutional Recurrent Neural Networks (CRNNs), and Convolutional Temporal Convolutional Networks (CTCNs), the paper highlights consistently low error rates, often below 2%, which align with top benchmarks in the literature. Additionally, DNNs outperform standard methods like Support Vector Regression (SVR) and Random Forest (RF), reducing error rates by over 50% in certain cases. The authors of [68] employed a sequential model with five hidden layers for gender detection. Using the trained model on the Common Voice dataset, they achieved an accuracy of approximately 91% for gender detection.

The study [39] explores handcrafted features such as Mel spectrogram, MFCC, Chroma, spectral contrast, and Tonnetz, analyzing their performance across various classifiers, including K-Nearest Neighbors (KNN) and Multilayer Perceptrons (MLP). Additionally, deep learning models such as Deep Neural Networks (DNN), ResNet34, and ResNet50 are evaluated for their performance on spectrogram images. Notably, ResNet50 emerges as the most effective model, achieving an impressive 98.57% accuracy on the Common Voice dataset while demonstrating strong generalization capabilities across different datasets. In their paper [69], the authors propose using DNNs to encode each utterance into a fixed-length vector by pooling the activations of the last hidden layer over time. The feature encoding process is designed to be jointly trained with an utterance-level classifier for improved performance. Experiments on a Mandarin dataset demonstrate the effectiveness of their proposed method for age and gender recognition tasks, achieving an accuracy of 92.72%. DNN-based embedder architectures, such as the d-vector system, have shown robust performance, with accuracies ranging from 96.8% to 99.6%, depending on the training and testing datasets. The highest result of 99.6% accuracy for gender recognition was achieved when the model was trained on the Common Voice dataset and fine-tuned on the TIMIT dataset [70].

While most studies focus on adult voices, some research has also investigated children's voices. Identifying gender in children is more challenging since sex-based differences in vocal tract structure do not develop until puberty [71]. Studies exploring gender detection in children's voices include [72–75].

Looking at the different works on gender detection in Table 2, it is evident that earlier efforts primarily focused on manually extracting features and feeding them into suitable machine learning models for classification. Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) were prominent choices, laying the foundation for the field's development. The introduction of deep learning techniques marked a significant shift, enabling the automatic extraction of intricate features with minimal manual intervention. A diverse set of deep learning architectures, such as Long Short-Term Memory networks (LSTMs) and Convolutional Neural Networks (CNNs), has been employed to improve accuracy. Multi-output models like 1D CNNs have proven effective in handling complex speech data. The adoption of

Table 2
Overview of the Algorithm, Features, and Datasets for gender detection.

Reference	Accuracy	Algorithm	Features	Dataset
Keyvanrad et al. [53]	96.53	SVM, GMM, MLP	MFCC + fused likelihood score	OGI Multilingual Corpus
Alsulaiman et al. [55]	96.44	Simpson's Rule	Peak Intensity	DARPA TIMIT
Bahari et al. [63]	96.00	WSNMF, GRNN	Gaussian mixture weight super vectors	Dutch Database
Djemili et al. [61]	96.40	GMM, MLP, VQ, LVQ	MFCC	IViE Corpus
Buyukyilmaz et al. [60]	96.74	MLP	22 acoustic features	Voice Gender Dataset
Pahwa & Aggarwal. [56]	93.48	SVM	MFCC + Delta + Delta-Delta	Vowel data
Wang & Tashev. [69]	92.72	DNNs	Energy + pitch + voice probability + 26 dim log MEL spectrograms	Mandarin dataset
Chaudhary & Sharma. [57]	96.45	SVM	MFCC + pitch + energy	Self data - 280 samples
Ertam et al. [65]	98.40	Deeper LSTM	IRQ + meanfun + sfm + sd + median + Q25 + Q75 + mode + centroid + meandom	Voice Gender Dataset
Uddin et al. [58]	96.80	SVM, KNN	FF + spectral frequency + spectral flatness + mode frequency + MFCC	TIMIT
Uddin et al. [59]	93.01	Multi-output 1d CNN	FF + spectral frequency + spectral flatness + mode frequency + MFCC + LPC	RAVDESS + TIMIT + BCG
Zaman et al. [76]	96.4	CatBoost, Random Forest, XGBoost, KNN, SVM	20 statistical feature	Common Voice Dataset
Kwasny et al. [70]	99.6	DNNs	d-vectors	TIMIT
Livieris et al. [66]	98.42	iCST-Voting	–	Voice Gender Dataset
Alnuaim et al. [39]	95.63	KNN, DNN, MLP	MFCC, Mel spectrogram, Chroma STFT, Tonnetz, special contrast	Common Voice Dataset
Alnuaim et al. [39]	97.94	ResNet34	Spectrograms	Common Voice Dataset
Alnuaim et al. [39]	98.57	ResNet50	Spectrograms	Common Voice Dataset
Kone et al. [68]	91.00	DNN	Spectrograms	Common Voice Dataset

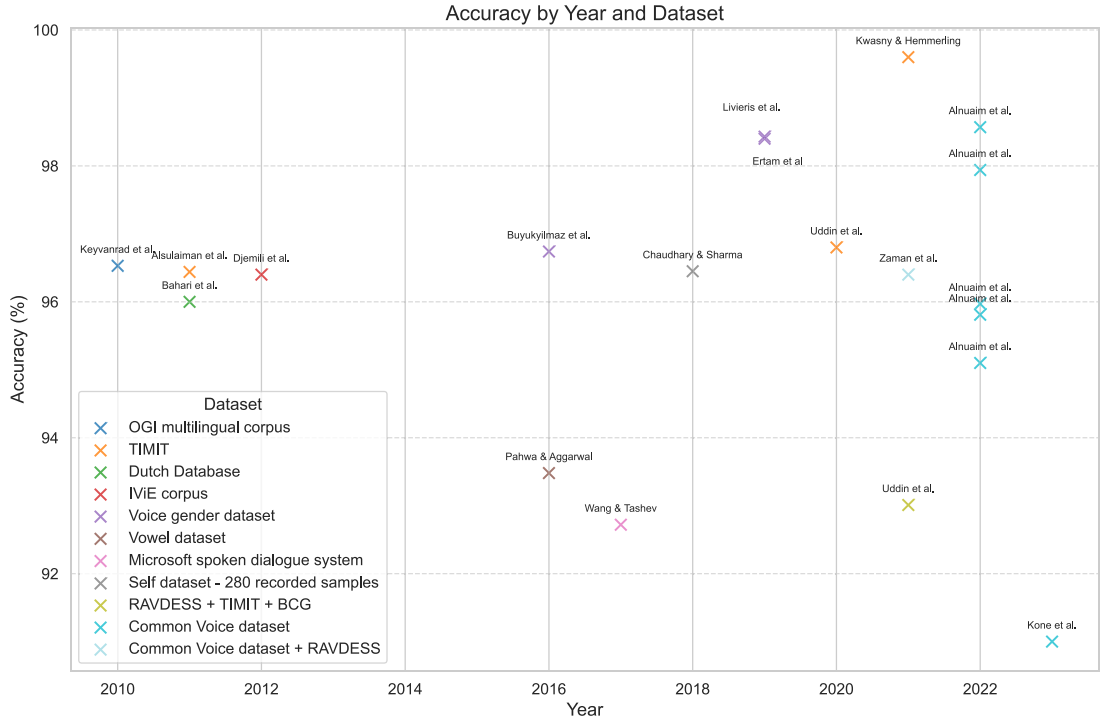


Fig. 6. Accuracy progression for gender detection across different datasets.

ensemble techniques, such as iCST-Voting, has demonstrated potential in boosting overall accuracy. Additionally, advanced architectures like ResNet34 and ResNet50, initially designed for computer vision tasks, have been adapted for gender detection, showcasing the transferability of algorithms across domains. Information on various datasets used for gender detection is presented in Table 7. A multidimensional comparison of accuracy across different datasets is illustrated in Fig. 6.

5.1.2. *Age inference from voice*

Age detection is a challenging task as several factors, such as the shape of the vocal tract, health, emotional state, gender, and accent, can influence speech. However, certain voice characteristics provide indications of a speaker's age. For instance, younger speakers generally exhibit a higher speech rate [77], and the fundamental frequency tends to decrease with age [78].

One of the pioneering studies for age estimation based on paralinguistic features like speech rate was conducted by authors of [79]. The authors used acoustic, not linguistic, information in their utterances. Earlier methods of age estimation relied upon using statistical features and models. For instance, authors of [80] conducted a study to predict age from speech using Hidden Markov Model (HMM) and SVM. The accuracy of the predictive model using HMM is 48.1% to 70.7%, and the accuracy using SVM is 53.1% to 72.6% on the AT&T's, HMIHY "how May I Help You", corpus [81]. In the paper [63], authors proposed a novel hybrid architecture combining WSNMF and GRNN. Evaluation on the Dutch dataset [64] shows that the MAE of age estimation using the proposed hybrid method is 7.48 years. Similarly, [82] conducted a study to predict age from speech using a two-level GMM into four categories: child, young, adult, and senior. The summary age classification of the whole GMM-based system was 92.3% on the Czech and Slovak database [83].

Recent approaches have focused on using machine learning and deep learning approaches with more inclination on the deep learning side. Deep learning approaches are adept to learning complex patterns in the data, and have shown better performance than the machine learning models. Machine learning approach as suggested by authors of the paper [68] propose a grid search pipeline technique using RobustScalar, Principle Component Analysis and Logistic Regression algorithms for age prediction on the common voice dataset and were able to achieve an accuracy of 59%. With a focus towards efficient feature extraction and deep learning based solutions, the authors of [84] introduce a novel embedding framework known as x-vector, which relies on a deep neural network (DNN). They conducted training on a time-delay neural network (TDNN) for a specific speaker classification task. In a similar context, the authors of [85] also propose an embedding approach called d-vector for speaker verification. Like the method detailed in [84], both approaches aim to transform variable-length utterances into fixed-size embedding vectors. However, these methods diverge in how they generate the embedding vector. In the x-vector system, statistical pooling is applied to the output of the final hidden layer of a convolutional neural network (CNN) to aggregate global context. In contrast, the d-vector architecture relies on a straightforward multi-layer, long-short time memory (LSTM) recurrent neural network (RNN), where the output of the last cell in the last hidden layer is utilized as the embedding.

Efforts have also been made to adapt both the x-vector and LSTM frameworks for age estimation tasks. In [86], the authors trained an LSTM-based system that demonstrated superior performance compared to the i-vector baseline when dealing with brief speech segments, specifically on the NIST SRE 2010 dataset [87]. Meanwhile, in another study documented in [41], a research group introduced an approach based on x-vectors, achieving a mean absolute error (MAE) of 4.92 years. Notably, when implementing the i-vector system on the same dataset, the MAE was higher at 5.82 years, indicating a clear advantage of the x-vector approach in this context. The paper [88] introduces a novel DNN architecture that can work with small training dataset for age prediction. The DNN system is able to improve the age RMSE by at least 0.6 years as compared to a traditional SVR system trained on GMM mean supervectors. The RMSE errors are 7.60 and 8.63 years for male and female respectively on the TIMIT dataset with an average speech duration of 2.5 s. Authors also claim at most 3% performance degradation with 1-s speech input compared to the whole duration.

Using a set of common features to estimate a number of physical traits, authors of the paper [88] trained a support vector regressor to achieve a MAE of 5.2 years for male and 5.6 years for female speakers on the TIMIT dataset. Several other traits including shoulder size, waist size, and weight have been analyzed as well [78]. The utilization of x-vectors and d-vectors, along with the application of transfer learning, has been explored in the simultaneous estimation of age and gender, as described in [70]. This research addresses the challenge of having limited training data by using transfer learning from networks pre-trained for tasks not directly related to speaker profiling, drawing on well-known speech recognition datasets such as TIMIT and common voice datasets. The application of transfer learning results in MAE of 5.12 years and 5.29 years, in addition, an RMSE of 7.24 and 8.12 years for male and female speakers respectively. Authors of the paper [89] highlight that majority of the previous works have predominantly relied on deep learning techniques applied to either hand-crafted features [88] or relied on complex structures involving millions of parameters [90,91]. These approaches have shown various limitations, such as the high cost associated with feature extraction and the dependence on manually selected features. These drawbacks may render them less suitable for real-time applications. Consequently, the authors propose an end-to-end speaker profiling system designed to estimate age, height, and gender. A novel **wavelet filter-bank initialization** method for CNN with residual blocks is proposed and tested on the TIMIT dataset and it achieved comparable results with more popular DNN with embeddings. The system achieved a MAE of 5.36 years and 6.07 years for male and female speakers in age estimation. A summary of the major works for age estimation from voice is shown in Table 3 and a list of datasets are included in Table 7.

5.1.3. Emotion recognition from voice

Emotions can be recognized through various means, including direct questioning, tracking implicit parameters or vital signs, voice recognition, facial expressions, and gesture recognition. However, Speech Emotion Recognition (SER) remains a complex field with ongoing research employing both machine learning and deep learning approaches. Several traditional machine learning methods have been applied to SER, including Gaussian Mixture Models (GMMs) [92], Hidden Markov Models (HMMs) [93], Support Vector Machines (SVMs) [94], and k-Nearest Neighbors (KNNs) [95]. These methods typically involve extracting a feature set from speech and then training a classifier. Bozkurt et al. [96] pioneered the use of Line Spectral Frequencies (LSFs) for emotion recognition, demonstrating improved classification rates compared to Mel-Frequency Cepstral Coefficients (MFCCs) when using a GMM classifier on the Berlin Emotional Speech Dataset (EMO-DB) [97] and the FAU Aibo Emotion Corpus [98].

Using the EmoSTAR dataset [99], Korkmaz et al. [100] investigated the use of Mel-Frequency Cepstral Coefficients (MFCCs) for analyzing emotional content in speech. MFCCs were extracted from spoken utterances using overlapping frames to ensure smooth transitions and minimize data loss. The impact of frame length and frame shift (also known as "scroll time") on emotion recognition performance was evaluated using Support Vector Machines (SVMs) and k-Nearest Neighbors (k-NNs). Their 10-fold cross-validation analysis yielded a notable accuracy of 98.7% in classifying emotions. The authors of [51] used MFCC features obtained from the EMO-DB dataset and used NN, SVM classifiers for emotion recognition with an accuracy of ranging from 73–85.8%. Motamed et al. [101] also leveraged MFCCs from the EMO-DB dataset but adopted a different approach, using an Adaptive Neuro-Fuzzy Inference System (ANFIS) and a Multilayer Perceptron (MLP) network inspired by the structure of the brain's emotional network. In their model, the ANFIS component, simulating the amygdala and orbitofrontal cortex, generated classification rules. These rules were then fed into the MLP network to classify speech emotion signals, achieving an accuracy of 72.5% on the EMO-DB dataset.

To reduce computational cost, Liu et al. [102] proposed a feature selection method based on correlation analysis and the Fisher criterion to remove highly correlated redundant features. Additionally, they introduced an emotion recognition approach using an extreme learning machine (ELM) decision tree, considering the confusion degree among different basic emotions to improve classification accuracy. Experiments on the CASIA Chinese speech database [103] demonstrated an average recognition rate of 89.9%.

Table 3
Overview of the Algorithm, Features, and Datasets for age detection.

Ref.	Year	Dataset	Algorithms for age	Features	Metrics for age (Accuracy %)MAE (years)& RMSE (years)
Shafran et al. [80]	2003	HMIHY	HMM,SVM	Voice Signatures	72.6%
Bahari et al. [63]	2011	Dutch Database	WSNMF,GRNN	Gaussian Mixture weight super vectors	MAE = 7.48 years
Prilbil et al. [82]	2016	Czech and Slovak	GMM	–	92.3%
Wang & Tashev [69]	2017	Mandarin dataset	–	Energy + Pitch + Voice Probability + 26 dimension Log MEL Spectrograms	92.72%
Zazo et al. [86]	2018	NIST SRE08 + SRE10	LSTM-RNNs	MFCC + Pitch + Probability of Voicing + Normalized Cross Correlation Function	MAE = 6.23 for male and 7.31 for female
Ghahremani et al. [41]	2018	NIST SRE08 + SRE10	x-vector DNN	23-dim MFCC short-time mean normalized over sliding the window of 3 s	MAE = 4.92 years
Kalluri et al. [88]	2019	TIMIT	DNNs	MFCC + Delta + Delta-Delta	RMSE = 7.6 for male and 8.63 for female
Kalluri et al. [78]	2020	TIMIT	SVR	Log-mel Spectrograms + Formant + Harmonic Features	MAE = 5.2 for male and 5.6 for female
Zaman et al. [76]	2021	Common Voice dataset + RAVDESS	CatBoost, Random Forest, XGBoost, KNN, SVM	20 statistical features	70.4%
Kwasny and Hemmerling [70]	2021	TIMIT	DNNs	d-vector feature extractor with front-end modules pre-training on Common Voice	RMSE = 7.24 for male and 8.12 for female
Jaid et al. [89]	2023	TIMIT	CNN	Residual Blocks + Filter Banks	MAE = 5.36 for male and 6.07 for female
Tursunov et al. [91]	2021	TIMIT	CNN	Multi-attention Module	73%
Kone et al. [68]	2023	Common Voice dataset	RobustScalar, PCA, and Logistic Regression	Spectrograms	59%

Deep learning methods have also demonstrated their superior performance compared with traditional machine learning methods in emotion recognition, primarily due to their ability to automatically learn complex features, scalability, and higher recognition accuracy [15, 104]. According to a comprehensive survey by Hashem et al. [15], DL techniques for emotion recognition can be categorized into three main approaches: (1) extracting handcrafted features followed by ML classification, (2) using DL for classification with either handcrafted features or automatically extracted features through DL layers, and (3) converting sound waves into spectrogram images to be used as input for DL models. While various methodologies have been employed in the field of SER to achieve acceptable results, the state-of-the-art performance is often achieved using DL techniques. Commonly used DL architectures in SER include CNNs and their variants, DBNs, Autoencoders, LSTMs, and RNNs.

The authors of the study [105] have used MFCC features and a 1D CNN architecture for SER, achieving 82.3% accuracy on the RAVDESS dataset for six emotion classes. Similarly, Issa et al. [106] proposed a 1D CNN-based approach that extracts MFCCs, chromagram, spectrograms, Tonnetz representations, and spectral contrast features for SER. Their model achieved speaker-independent classification accuracies of 76.1%, 86.1%, and 64.3% on the RAVDESS, Berlin EMO-DB, and IEMOCAP datasets, respectively. An improvement over this technique was proposed by the authors of [107]. Instead of examining the whole utterance to recognize the final state of emotion, the authors used key sequence segment selection based on Radial-Based Function Network (RBFN) similarity measurement in clusters to recognize spatial–temporal information while reducing complexity. In addition, the normalized CNN features are fed to the deep BiLSTM to learn the temporal information for recognizing the final state of emotion. Experiments over the IEMOCAP, EMO-DB and RAVDESS datasets resulted in accuracies of 72.25%, 85.57% and 77.02% respectively.

CNNs effectively extract features from speech spectrograms [108], while fully convolutional networks (FCNs) are designed for dense prediction tasks but struggle with temporal modeling [109]. In contrast, RNNs and LSTMs excel at capturing temporal dependencies and are widely used for SER. To model spatio-temporal features, hybrid architectures combining CNNs with RNNs or LSTMs are commonly employed [15,104]. Zhao et al. [110] proposed a model integrating attention-based bidirectional LSTMs with attention-based fully convolutional networks to enhance spatio-temporal feature learning in SER. Their approach achieved a weighted accuracy (WA) of 68.1% and an unweighted accuracy (UA) of 67.0% on IEMOCAP, along with 45.4% UA on FAU Aibo Emotion Corpus (FAU-AEC). Deep representations, when combined with a linear support vector classifier, performed comparably to standard acoustic feature sets such as extended Geneva Minimalistic Acoustic Parameter Set **(eGeMAPS)** [111] and Computational Paralinguistics Challenge features set **(ComParE)** [112]. Anvarjon et al. [113] addressed two key challenges in SER: reducing computational complexity and improving accuracy. They proposed a lightweight CNN with plain rectangular kernels and a modified pooling strategy, emphasizing deep frequency features in speech spectrograms. Their model achieved recognition accuracies of 77.01% on IEMOCAP and 92.01% on EMO-DB, surpassing state-of-the-art SER models.

Mustaqeem et al. [114] designed a one-dimensional dilated CNN architecture for real-time SER, similar to Zhao et al. [110]. Their approach combined CNNs for local feature extraction and LSTMs for global feature learning using a simple strategy, achieving an unweighted accuracy of 67% on the IEMOCAP dataset. To overcome the limitations of [110], they proposed a multi-learning strategy framework incorporating two learning modules: a residual block with skip connections (RBSC) to capture emotional cues and a sequence learning (Seq_L) module to model long-term contextual dependencies. The proposed model achieved recognition accuracies of 73% on IEMOCAP and 90% on EMO-DB. Mishra et al. [115] introduced a deep learning framework that replaces MFCCs with log Mel-frequency spectral coefficients (MFSCs), which have been shown to outperform MFCCs. Their approach combines deep CNNs and BiLSTMs in an ensemble model. Experiments on the TESS [116] and SAVEE [117] datasets demonstrated an accuracy of 96.36%, surpassing single-model learners and previous machine learning techniques.

Apart from CNNs, DBNs [118] have also been explored for automatic SER tasks. Deep learning techniques inherently capture complex non-linear features in multimodal data. Kim et al. [119] demonstrated the effectiveness of deep learning for emotion classification, testing a suite of DBN models on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [120], achieving an average success rate of 73%. Xia et al. [121] proposed a framework combining DBNs with

Table 4
Overview of the classifiers, features, datasets, recognized emotions, and main contributions for SER.

Ref.	Year	Algorithm	Features	Dataset used	Recognized emotion	Main contribution
[51]	2015	NN, SVM	MFCC	EMO-DB	Anger, Fear, Joy, Sad, Neutral, Disgust, Boredom	Present a novel architecture based on NN for SER improving classification rate in contrast to traditional algorithms
[101]	2017	ANFIS, MLP	MFCC	EMO-DB	Anger, Fear, Joy, Sad, Neutral, Disgust, Boredom	Emotional learning models with different structures and functions using the amygdala-orbitofrontal cortex subsystem are suggested
[102]	2018	ELM DT, SVM	Prosodic features	CASIA	Happy, Sadness, Surprise, Angry, Fear, Neutral	A novel feature selection method based on correlation analysis and Fisher is introduced that eliminates redundant features
[123]	2018	Adversarial auto-encoders	Spectral, prosody and energy-based features	IEMOCAP	Happy, Sad, Anger, Neutral	Applied auto-encoders to enhance the state of research in emotion recognition
[106]	2020	DEEP CNN	MFCC, Spectrogram, Chroma, Spectral Contrast, Tonnetz	RAVDESS, IEMOCAP, EMO-DB	Happy, Sad, Anger, Calm, Fear, Nervous	All the models can work directly on raw sound data without conversion to other visual representations, and the approach sets a new state-of-the art on RAVDESS and IEMOCAP datasets.
[107]	2020	CNN BiLSTM	Spectrograms	RAVDESS, IEMOCAP, EMO-DB	Happy, Sad, Anger, Calm, Fear, Nervous	A key sequence segment selection based on radial based function network (RBFN) similarity measurement in clusters is proposed that recognizes the Spatio-temporal information while reducing complexity
[105]	2021	1D CNN	MFCC, Pitch	RAVDESS	Happy, Sad, Anger, Calm, Fear, Nervous	Improved SER rate using MFCC features on RAVDESS dataset
[115]	2023	Deep CNN + BiLSTM	MFSC	TESS+SAVEE	Happy, Sad, Anger, Neutral, Fear, Surprise, Disgust	Used MFSC feature and trained a novel ensemble deep learning framework that outperforms conventional machine learning techniques

i-vector space modeling for acoustic emotion recognition. Using MFCC features, their model was evaluated on the IEMOCAP dataset and achieved an accuracy of 59.6%. Autoencoders [122] have also been investigated for emotion recognition. Sahu et al. [123] explored their ability to transform high-dimensional feature vectors into a compressed space while preserving emotion-class distinguishability. Additionally, they examined how synthetic sample generation in the original feature space could aid in training emotion recognition classifiers. Experiments on the IEMOCAP dataset yielded a success rate of 57.88%. An overview of studies in emotion recognition is provided in Table 4 and Table 7 also highlights the most commonly used ones in the studies covered by this survey.

While this survey primarily focuses on audio-based profiling methods, it is important to acknowledge that human emotion perception is inherently multimodal. In actual human interactions, emotions are expressed not just through speech but also via facial expressions, hand movements, and physiological signals. Consequently, hybrid approaches that incorporate multiple modalities have been explored to improve emotion recognition performance. Several studies have investigated emotion recognition using facial expressions [124], hand gestures [125], body movements [126], and electroencephalography (EEG) signals [127]. A comprehensive survey on multimodal emotion recognition is available in [128].

The success of both unimodal and multimodal approaches is largely dependent on the availability of high-quality datasets. As machine learning and deep learning research continue to advance, a variety of datasets have been developed for emotion recognition. Table 7 highlights the most commonly used ones in the studies covered by this survey. Beyond individual datasets, various challenges and competitions, such as the Audio/Visual Emotion Challenge (AVEC) [129–131], have significantly contributed to research progress in this field. These challenges often introduce novel multimodal datasets, such as the SEWA dataset [132], and define specific tasks for affective state detection, cultivating innovation and enriching available resources.

5.1.4. *Mental health inference from voice*

Voice has been investigated for inference of Mental Health of a person such as depression, autism, schizophrenia. Authors of [133] conducted a study and found that voice acoustic features extracted from read speech demonstrated variable effectiveness in predicting clinical depression scores in men and women. Voice features are highly predictive of **HAMD (Hamilton Depression Rating Scale)** [134] and **BDI-II (Back Depression Inventory)** [135] scores which are tools to measure a patient's degree of depression and suicidal risk. The studies [136,137] point out the peculiar speech patterns associated with depression includes decreases in intensity, speech rate, stress, loudness, sluggishness in articulation, monotony and lack in vitality. The study [138] investigates the role of glottal features in speech for diagnosing clinical depression. It combines glottal with prosodic and vocal tract features for analysis. Feature selection strategies were tested across different domains, revealing that **glottal and prosodic features** together offer better discrimination of depression in speech than other combinations. Quadratic discriminant analysis was used for classification, highlighting the importance of glottal features in the assessment of depressed speech.

Authors of [139] investigate voice as a potential biomarker for depression suicidality, **psychomotor disturbance**. Data was collected from web version of PHQ9 (Patient Health Questionnaire) hosted by Mental Health America (MHA). A score of AUC of 0.821 and a MAE of 4.7 was obtained indicating voice as a potential biomarker. It is possible to identify high risk adolescents of early depression even as early as two years before as stated by the authors of [140]. Using a novel multi-classification approach aided by Glottal [141], Prosodic, Teager energy operator (TEO) and Spectral features [142,143] and a weighted classification decision procedure, an accuracy of 73% and a desirable sensitivity-to-specificity ratio of 79%/67% was obtained.

The study [5] showcases voice as a non-evasive and potential biomarker for the early detection of **depression**. Upon investigation using the EEG (Electroencephalogram) and audio data from clinically depressed patients and matching normal controls, the study acquired

an accuracy of 83.4% on the MODMA (Multi-modal Open Dataset for Mental-disorder Analysis) dataset [144]. Similarly, authors of the paper [145] introduced a multi-model approach to enrich detection of mental disorders like depression using text and learning from low level and graph-based voice signal features. The study improves accuracy employing a transformer-based deep learning architecture that uses novel graph-based features on the DAIC-WOZ (Distress Analysis Interview Corpus/Wizard-of-Oz set) dataset [146].

Along with depression, insights into the identification of **Schizophrenia** based on the voice have been studied. This study [147] focuses on comparing the speech production fluency of patients with schizophrenia to healthy controls using a temporal speech parameter set. The results show that the temporal indicators can capture specific differences in spontaneous speech, with classification accuracy scores between 70%–80% and F-measure scores between 81% and 87%. The pause-related temporal parameters that consider both silent and filled pauses were found to be the most useful for distinguishing the two speaker groups.

A study [148] highlighted the potential of using voice characteristics for detecting autism. The study found that children with **ASD (Autism Spectrum Disorder)** often exhibit atypical patterns in prosodic elements such as monotonous pitch, reduced stress, odd rhythm, flat intonation, and even differences in the harmonic structure of their speech. These are among the earliest signs of the disorder. Another study [149] compared the acoustic features of the speech of TD (Typically Developing) children and children with ASD. It found that for all children with ASD, voice and speech are characterized by high values of pitch, an abnormal spectrum, and well-marked high-frequency. Another study [150] states that easy-to-measure voice acoustic parameters can be used as a diagnostic aid tool, specific to ASD. The model achieved an overall accuracy of 91% against TD (Typically Developing) children. Thus voice can potentially contribute significantly to the early detection and intervention of ASD.

Various studies have explored the potential of voice as a marker for **Parkinson's disease** (PD), a neurodegenerative disorder characterized by motor deficits such as bradykinesia (slow movement), rigidity, and tremors [151]. PD is estimated to be the second most common degenerative disorder, after Alzheimer's, and affects approximately 12 individuals per 1000 of the population with an increased prevalence in persons aged over 65 [152]. Authors of the study [153] employ a method that separates speech into voiced and unvoiced segments and models the energy content of unvoiced sounds using Mel-frequency cepstral coefficients (MFCCs) and Bark band energies (BBEs). The method was tested on dataset [154] that contains different speech tasks performed by speakers in Spanish, German, and Czech. It outperformed classical approaches, achieving accuracies ranging from 85% to 95% in classifying speech of people with PD and healthy controls. The method can detect PD in early and middle stages of the disease and is robust against different technical conditions. It shows promise for future development of computer-aided tools for the automatic evaluation of dysarthric speech signals. Datasets for mental health detection are listed in Table 7. Furthermore, specific datasets to model health outcomes can be found here.[1]

Challenges and competitions such as Audio/Visual Emotion Challenge **(AVEC)** have carried out competitions focusing on mental health issues such as bipolar disorder recognition [155], depression detection [156], state-of-the-mind recognition [157]. These challenges introduce multimodal datasets that focus on detecting specific mental health conditions.

[1] Available at: https://github.com/talhanai/speech-nlp-datasets.

5.1.5. Inference of voice pathology

Voice pathology detection involves labeling phonation as either normophonic or dysphonic, as described in [158]. Various medical techniques, such as laryngoscopy, glottography, stroboscopy, electromyography, and videokymography, are available for directly examining and diagnosing pathologies. However, these methods have drawbacks: the human vocal tract is difficult to access during phonation, which hinders accurate identification of pathologies. Additionally, these diagnostic methods can be uncomfortable for patients and may distort the signal, leading to incorrect diagnoses [159]. Consequently, non-invasive assessment methods are gaining popularity due to their robustness, low cost, comfort, and reduction in subjective bias. For voice pathology detection, the most commonly used datasets include the **Saarbruecken Voice Database (SVD)**, **Arabic Voice Pathology Database (AVPD)**, and **Massachusetts Eye and Ear Infirmary Database (MEEI)**. Details about these datasets are presented in Table 7.

Voice pathology detection remains an active research area with numerous studies exploring innovative methods. The authors of [160] proposed a CNN–LSTM architecture that operates directly on raw audio signals, eliminating the need for pre-extracted feature vectors. Their experiments on the SVD dataset, using voice recordings of sustained vowel /a/ at normal pitch, achieved 71.36% validation accuracy and 68.08% test accuracy—results comparable to earlier studies employing different methodologies. Another notable study [161] introduced CNN-based pathology classification, leveraging a pre-trained model fine-tuned to extract features from voice signals. This approach achieved an impressive 95.41% accuracy, with F1-Score and Recall values of 94.22% and 96.13%, respectively. The system demonstrates strong potential for real-world clinical applications, enabling fast, automatic diagnosis within 3 s. However, both studies face limitations, such as the small size of the SVD testing dataset, the lack of gender-specific analyses, and insufficient consideration of pathology severity. Addressing these challenges is essential for improving the reliability and scalability of voice pathology detection systems.

A novel approach to utilizing voice signals for the classification and detection of Upper Respiratory Tract Infections (URTI) was proposed by the authors of [162]. The study evaluated the efficacy of a classifier in detecting speech affected by URTI and demonstrated its ability to achieve results comparable to those obtained in related health-based detection tasks, such as autism detection [163], Parkinson's disease detection [164], and cognitive or physical load classification [165]. These findings highlight the potential of computational paralinguistic analysis for detecting URTI-related illnesses. The work leverages the dataset described in [166]. Additionally, Amazon Technology Inc. has patented a system [24] that identifies pathological voices and seamlessly integrates this knowledge to provide tailored responses and suggestions.

In addition to these, several studies have explored different learning models for voice pathology detection. These include stochastic models, such as GMM-based approaches [167–169]; machine learning models, like SVM-based methods [170,171]; and deep learning models [160, 172], as summarized in [173]. While recent studies primarily focus on differentiating normal and pathological voices, they often struggle with fine-grained classification of specific conditions, such as laryngeal cancer or voice tremors. Current machine learning approaches face challenges related to generalizability, dataset imbalance, and robustness to variability in speech signals. To address these limitations, future research should focus on developing more discriminative feature extraction techniques that capture subtle pathological variations in speech while enhancing classifier architectures with better regularization, improved activation functions, and domain-adaptive learning strategies. These advancements would improve accuracy, efficiency, and the overall reliability of voice pathology detection systems.

5.1.6. Inference of personality traits

A person's personality is a blend of their behavior, emotions, motivations, and unique patterns in their thinking. It significantly influences life decisions, well-being, health, and shaping preferences and aspirations. As a result, the capacity to automatically identify someone's personality traits holds immense practical significance across various domains.

Various personality modeling scales have been used by the researchers, such as Eysenck Personality Questionnaire-Revised (EPQ-R) and the Eysenck Personality Profiler (EPP) [174], Sixteen Personality Factor (16PF) [175], and the three-trait PEN (Psychoticism, Extraversion, and Neuroticism) model [176]. Furthermore, Myers-Briggs Type Indicator (MBTI) [177] also categorizes people into four dimensions: sensing/intuiting, thinking/feeling, judging/perceiving, introversion/extraversion. However, in the realm of automated personality inference, the prominent model is the "Big Five" model [178]. These traits are based on binary (yes/no values):

1. **Openness (O)**: Is the individual inventive and curious or dogmatic and cautious?
2. **Conscientiousness (C)**: Is the individual efficient and organized or sloppy and careless?
3. **Extraversion (E)**: Is the individual outgoing, talkative, and energetic or reserved and solitary?
4. **Agreeableness (A)**: Is the individual trustworthy, straightforward, generous, and modest or unreliable, complicated, meager, and boastful?
5. **Neuroticism (N)**: Is the individual sensitive and nervous or secure and confident?

There have not been many comprehensive literature reviews on personality detection. Additionally, audio-based datasets are rare. Some of the audio-based dataset for personality detection is presented in Table 7. Personality detection is a new and upcoming field. There have been studies for personality detection based on text, and visual features, and relatively few studies [179–181] have only used audio as the sole input. A study [179] describes an automated system for speaker-independent personality prediction in the context of human-human conversations using the PersIA (**P**ersonable and **I**ntelligent virtual **A**gents) speech dialog corpus annoted with user self-assessments of the Big-Five personality traits. Study claims to have promising results on detecting the conscientiousness and extroversion labels.

Another study [181] examines inferring speakers' personality traits from spontaneous conversations, using the Big-Five model. It outlines annotation methods applied to 128 speakers from the AMI corpus, followed by experiments using various features. Findings show accurate recognition of high/low extraversion, conscientiousness, and neuroticism, but not agreeableness and openness. Non-linguistic features outshine linguistic factors in this analysis.

For automatic extraction of personality traits the authors of [182] use the **NEO-FFI inventory**, which measures the Big-Five personality traits. The authors analyze the ratings and find that they are generally consistent and correlate well with the speaker's instructions. They also find that some factors, such as openness and extroversion, are more difficult to manipulate or perceive. The authors extract various acoustic and prosodic features from the speech recordings and use support vector machines to classify them into 10 classes. They achieve about 60% accuracy, which is much higher than chance level. They also find that some features such as MFCCs, intensity, and pitch, are more informative than others. However, limitations such as database having a single professional speaker, unreliable link between the trained professional and actual untrained speakers etc are the downside of the approach. Often times a multi-model that works using text, visual features, and audio are more relevant in personality detection.

Text is one of the most common and rich sources of information for personality detection, as it reflects the style, content, and sentiment of the writer or speaker. Linguistic information has been widely used because self-assessment and peer-assessment of personality leverage a list of verbal descriptors. Subsequently, these lists have been combined and condensed into higher level dimensions. Authors of [183] demonstrate that certain acoustic and linguistic features can be indicative of perceived leadership qualities, such as charisma, decisiveness, and confidence. Text-based personality detection can be done using closed-vocabulary or open-vocabulary methods. **Closed-vocabulary** methods rely on predefined categories of words, such as LIWC (Linguistic Inquiry and Word Count), MRC (Medical Research Council), or Mairesse, which are associated with psychological constructs. **Open-vocabulary** methods rely on extracting a comprehensive collection of language features from text, such as n-grams, punctuation, emoticons, topics, or word embeddings. Furthermore, deep learning models, such as CNNs, RNNs, LSTMs, GRUs, or bi-directional variants can be employed for personality detection [184]. These models can learn hierarchical, vector, and temporal representations of words and sentences, and can capture complex and subtle patterns in language use. Text-based personality detection faces several challenges, such as dealing with noisy, sparse, or unstructured data, modeling inter-trait dependencies, accounting for cultural and contextual variations, ensuring fairness and ethics, and explaining the results [185].

Incorporation of more than one source has also been carried out in personality prediction, the primary reason being the limited capability of single modality personality trait recognition. Bimodal architectures that combine fusion of features from the audio and visual modalities [186] and trimodal architectures that combine audio, video and textual modalities [187] are showing promising result. [188] introduced a multimodal deep learning approach that combined raw audio and visual elements to predict personality traits. They utilized a 14-layer 1D CNN for audio feature extraction and a pre-trained ResNet-50 network for visual feature extraction. Their method incorporated a fully connected layer to collectively learn audio-visual features for recognizing personality traits, achieving an average score of 91.6%.

On the other hand, [189] proposed a multimodal personality trait recognition technique incorporating audio, visual, and text components. They employed a ResNet-18 for extracting audio and visual features, while leveraging the skip-thought vectors for text features. Their approach involved a late fusion strategy to merge all three modalities, resulting in an average score of 91.61%. Similarly, the research [190] introduces a multimodal approach for recognizing personality traits, employing a combination of CNN, Bi-LSTM, and Transformer networks. The fusion of these techniques aims to grasp comprehensive audio-visual spatio-temporal features crucial for identifying personality traits. Additionally, the study conducts a comparative analysis of multimodal personality prediction utilizing three fusion methods. Through experiments conducted on the ChaLearn First Impression-V2 dataset, it was determined that decision-level fusion yielded the most superior results for multimodal personality trait recognition, achieving average score of 91.67%. A more detailed inspection on the efficient fusion (Feature-level, Model-level, Decision-level) of the multimodal is needed. A consolidated comparison of major studies employing the Big-5 personality measure—across unimodal and multimodal setups—is presented in Table 5.

A new direction in the field of personality recognition by offering an ethical and privacy-preserving methodology distinct from traditional approaches is suggested by the authors of [12]. While promising, the method to extract personality traits from conversational dynamics is relatively new, and its efficacy across broader and more diverse datasets cannot be fully guaranteed at this stage.

5.1.7. Multiple personal attributes

Numerous studies have been conducted in this field, ranging from investigations into singular attribute detection to research focusing on the simultaneous detection of multiple attributes from a unified feature set. In this section, we highlight studies that focus on inferring multiple attributes, including combinations such as age and gender, gender and

Table 5
Overview of the Modalities, Accuracy, Features and Algorithm for Personality Inference using BIG-5 Measure.

Ref.	Modality	Accuracy (%)	Features & Algorithm	Year
[190]	Audio, Video	91.67	VGGish model for extracting audio and VGG-Face model for visual features extraction	2023
[189]	Audio, Video, Text	91.61	ResNet-18 for extracting audio and visual features, while leveraging skip-thought vectors for text features.	2022
[188]	Audio, Video	91.6	14-layer 1D CNN for audio feature extraction and a pre-trained ResNet-50 network for visual feature extraction.	2021
[186]	Audio, Video	90–91	Facial features extracted with ResNet34 and audio features are combined and fed to Deep LSTM	2017
[187]	Audio, Video, Text	89.18	Facial features are extracted using OpenFace [191], Audio features such as MFCC, ZCR, spectral energy distribution, speaking rate are extracted using OpenSMILE and SenticNet is used for polarity detection from text.	2017
[184]	Text	57.99	1D convolutions to extract n-grams combined with Mairesse features	2017
[181]	Audio	57.48	prosodic features, speech activity features, word n-grams and dialog act tags were extracted and fed into simple ML classifiers	2012
[179]	Text	57.48	openSMILE feature extractor and a boosting based system for text categorization	2011
[182]	Audio	–	Cepstral features such as MFCC, ZCR, intensity, pitch, loudness, formants were fed into a SVM regressor	2010

Table 6
Overview of Feature extraction, Learning model and Audio profiling tasks.

Ref	Feature extraction		Learning model			Audio profiling tasks							Year
	Domain	Method	Statistical	ML	DL	Age	Gender	Emotions	MH	Personality	VP	ASC/AED	
[192]	FD	PRAAT		RF regres.		✔	✖	✖	✖	✖	✖	✖	2023
[193]	TFD	Statistical			MLP	✔	✔	✖	✖	✖	✖	✖	2020
[194]	TFD	Statistical		LSSVR		✔	✖	✖	✖	✖	✖	✖	2014
[195]	TFD	Statistical			ANN	✔	✖	✖	✖	✖	✖	✖	2015
[49]	TFD	Statistical		SVM		✔	✖	✖	✖	✖	✖	✖	2008
[4]	TFD	Statistical			MLP	✖	✔	✖	✖	✖	✖	✖	2020
[70]	TFD	NN Embeddings			DNN	✔	✔	✖	✖	✖	✖	✖	2021
[39]	TFD	Visual			DNN	✖	✔	✖	✖	✖	✖	✖	2022
[96]	TFD	Statistical	GMM			✖	✖	✔	✖	✖	✖	✖	2010
[100]	Cepstrum	Statistical		SVM/KNN		✖	✖	✔	✖	✖	✖	✖	2015
[196]	Cepstrum	Statistical	HMM			✖	✖	✔	✖	✖	✖	✖	2015
[197]	TD/Cepstrum	Statistical		SVM		✖	✖	✔	✖	✖	✖	✖	2011
[40]	TFD	Visual			CNN	✖	✖	✔	✖	✖	✖	✖	2019
[198]	TD	Statistical	GMM			✖	✖	✖	✔	✖	✖	✖	2003
[199]	TD+FD	Statistical		SVM/RF		✖	✖	✖	✔	✖	✖	✖	2016
[200]	TFD(CWT)	Statistical		SVM		✖	✖	✖	✖	✖	✔	✖	2007
[201]	Cepstrum	Statistical	GMM			✖	✖	✖	✖	✖	✔	✖	2009
[160]	TFD	NN Embeddings			DNN	✖	✖	✖	✖	✖	✔	✖	2017
[202]	TFD	Statistical		MC-SVM	1D_CNN	✖	✖	✖	✖	✖	✖	✔	2019
[203]	TFD	NN Embeddings			CNN	✖	✖	✖	✖	✖	✖	✔	2019
[204]	TFD	Opensmile		Boostexter		✖	✖	✖	✖	✔	✖	✖	2011
[205]	TD/FD	Statistical		Boostexter		✖	✖	✖	✖	✔	✖	✖	2012
[76]	FD	Statistical		MMLM		✔	✔	✔	✖	✖	✖	✖	2021

Notes:

- ✖- Not included; ✔- Included.
- TD - Time Domain FD - Frequency Domain, TFD - Time–Frequency Domain.
- ML - Machine Learning, DL - Deep Learning, ANN - Artificial Neural Network, DNN - Deep Neural Network, CNN - Convolutional Neural Network, MMLM - Multiple Machine Learning Models.
- SVM - Support Vector Machine, RF - Random Forest, MLP - Multilayer Perceptron.
- MH - Mental Health, VP - Voice Pathology, ASC/AED - Acoustic Scene Classification/Audio Event Detection.
- PRAAT - Phonetic and acoustic analysis toolkit.
- openSMILE - open-source Speech and Music Interpretation by Large-space Extraction.

emotions, and age and height. Several studies focus on the simultaneous inference of age and gender [70,206–209], whereas others examine age and height [42,88]. Speaker height estimation is grounded in the positive correlation between the size of the **Vocal Tract Length (VTL)** and a person's **height and weight** [71]. Studies focusing on weight estimation from voice remain scarce due to the lack of comprehensive datasets. Research has also explored the simultaneous inference of age, height, and gender [210,211]. Additionally, some studies have investigated predicting speaker body parameters, such as shoulder and waist size, based on voice characteristics [78,212,213]. Emotion recognition has been studied alongside age and gender, particularly in spoken dialog systems. For instance, [69] explores jointly recognizing these attributes using a deep neural network approach, demonstrating improved performance in Mandarin speech. Studies employing frequency spectrum analysis [214] suggest that speech signals can be used to simultaneously infer gender, age, and emotion. Notably, [76] demonstrates the feasibility of inferring all three attributes—age, gender, and emotion—from a single source for the first time. Additional details on the inference of multiple attributes can be found in Table 6.

5.2. Environmental attributes inference

The environment of a speaker can reveal sensitive information about their behavior, ethnicity, religion, choices, and more, which could potentially be misused for targeted marketing or other unethical purposes. This section explores current research on environmental awareness that can be leveraged for audio profiling.

Acoustic scene classification and event detection. Acoustic Scene Classification (ASC) is a critical task in the field of environmental sound analysis, aiming to categorize audio recordings based on the specific

environment in which they were captured [215]. ASC has garnered significant attention within the Audio and Acoustic Signal Processing (AASP) community. Voice recordings often contain ambient noise in addition to speech. By analyzing these background sounds, it is possible to gain insights into the environment where the audio was recorded. Examples include indoor environments (e.g., cafes, offices, libraries, grocery stores), outdoor environments (e.g., parks, pedestrian streets, city centers, urban parks, residential areas, forests, subways, metro stations), and modes of transport (e.g., buses, cars, trains, trams) [216, 217]. A challenging aspect of ASC is the detection of audio events that are temporarily present within an acoustic scene. Examples of such events include footsteps, sirens, doorbells, running water, and fire alarms. This task is referred to as Acoustic Event Detection (AED). Unlike ASC, which focuses on categorizing broader environmental contexts, AED emphasizes the precise temporal detection of specific sound events.

Numerous studies have been conducted on ASC/AED. For instance, the authors of [218] provide an in-depth overview of traditional feature extraction and classification techniques prior to the dominance of deep learning-based methods. Other surveys focus on state-of-the-art deep learning approaches for Acoustic Event Detection (AED) [9,219] and Acoustic Scene Classification (ASC) [10]. Additionally, several papers summarize algorithms presented in various ASC/AED challenges, such as [220].

ASC has garnered substantial attention in recent years, leading to the development of various state-of-the-art methods and the organization of challenges such as **CLEAR** [221] and the **DCASE (Detection and Classification of Acoustic Scenes and Events)** series, which began in 2013. These challenges have attracted participants from both academic and industrial backgrounds [220]. The DCASE challenges [222] serve as a platform for researchers specializing in the computational analysis of sound events and scene analysis to showcase and discuss their research outcomes. The classification of acoustic scenes involves the recognition of semantic entities, referred to as acoustic scenes, which is achieved through computational algorithms encompassing signal processing and machine learning techniques [223]. The **Interspeech Computational Paralinguistics Challenge (ComParE)** serves as a pivotal platform, driving innovation and advancements in Acoustic Event Detection (AED). Across its editions, ComParE has introduced a wide range of tasks specifically tailored to the detection, classification, and analysis of diverse acoustic events. These include:

- **Environmental Monitoring:** Detection of sounds such as mosquitoes [224].
- **Healthcare and Wellness Monitoring:** Detection of breathing, snoring, and heartbeats [166].
- **Pediatrics and Childcare:** Detection of baby sounds and infant crying [225].

Additional tasks from the ComParE challenge can be accessed here.

The use of deep learning-based methods has been a prominent area of research in ASC, with studies focusing on techniques such as **knowledge distillation, attentive max feature maps, weight quantization, and structured pruning** to enhance classification accuracy. Additionally, the development of robust and low-complexity ASC systems has been an area of interest, with models such as the ASC baseline, which features an inception-based and low-footprint architecture. The study by [226] investigates ASC for identifying audio recording scenes. It introduces a low-footprint ASC baseline model and compares it against other architectures. The study further enhances the baseline with a novel deep neural network (DNN), evaluating the trade-off between complexity and accuracy. It also examines the influence of sound events on ASC accuracy and proposes a method to integrate scene and event information. Experiments conducted on diverse datasets from the DCASE challenge demonstrate the models' effectiveness in real-world applications, particularly on edge devices and mobile platforms, emphasizing low computational complexity and general applicability. Furthermore, the paper introduces a visualization method to contextualize sound scenes.

In addition to acoustic scene classification, it has been demonstrated that specific acoustic events can also be recognized. These include various sounds related to animals (e.g., cat, cow, dog), natural soundscapes and water sounds (e.g., rain, wind, sea waves, crackling fire), non-speech human sounds (e.g., sneezing, drinking, clapping), domestic or interior sounds (e.g., door, can opening, mouse click, appliance sounds), and exterior sounds (e.g., siren, airplane, hand saw, church bells) [227,228].

Acoustic scene classification (ASC) and event detection (AED) are multifaceted tasks that have experienced significant advancements in recent years, particularly with the rise of deep learning-based methods and the organization of research challenges such as DCASE and ComParE. However, several challenges remain. High-performing ASC models are often computationally complex, making it difficult to optimize them for deployment on resource-limited devices such as IoT platforms. Additionally, the performance of ASC models lags behind other audio processing fields due to difficulties in extracting efficient features from diverse acoustic scenes. Furthermore, the availability of large-scale, inclusive datasets and a well-defined ontology remains a critical requirement. Table 7 presents some of the key datasets available for ASC/AED. Defining the boundaries of acoustic scenes is another challenge, as the definition of events and scenes often depends on specific use-case scenarios, adding to the complexity. Despite these challenges, the future of ASC/AED is promising. A concerted effort from the research community is needed to address these limitations and pave the way for more robust, efficient, and adaptable solutions.

6. Datasets

Audio profiling represents a multidisciplinary research domain encompassing diverse analytical tasks such as gender recognition, age prediction, emotion detection, accent recognition, environmental context detection and speaker characterization. The domain of audio profiling research has been fundamentally shaped by a select group of transformative datasets that have pushed the boundaries of acoustic analysis and machine learning technologies. These datasets represent more than mere collections of audio samples; they are critical infrastructures that have enabled breakthrough innovations in understanding human vocal characteristics. This section presents an overview of the most prevalent datasets, providing a comprehensive perspective on the dataset landscape.

Table 7 presents datasets encompassing diverse domains, modalities, and applications, illustrating the progression of audio profiling research. Prominent datasets such as TIMIT [229] and Common Voice [231] are characterized by their extensive speaker diversity and language coverage. TIMIT, a widely regarded benchmark for phoneme segmentation, is noted for its high-quality recordings and balanced dialects, making it indispensable for speech recognition and phonetic analysis tasks. In contrast, Common Voice excels in multilingual and demographic diversity, offering contributions from a global user base. This dataset is particularly suitable for building robust models for gender, age, and accent classification due to its wide-ranging representation.

For emotion-focused profiling, datasets such as IEMOCAP [120] and RAVDESS [236] are highly valued. IEMOCAP is praised for its nuanced emotional expressions and multimodal setup, but its relatively limited number of speakers restricts its generalizability across diverse populations. Similarly, RAVDESS, known for its controlled and validated emotional expressions, offers excellent quality but shares the limitation of restricted speaker variety, which may affect its applicability in real-world scenarios.

Gender and age profiling are addressed in datasets like the Dutch Corpus [64] and NIST SRE08 [230], both of which include balanced demographic distributions and metadata on speaker attributes. Large-scale resources, such as AudioSet [245] and Freesound Dataset 50k

Table 7
Datasets for audio profiling.

Dataset	Composition	Characteristics	AP tasks	Year
TIMIT [229]	630 speakers(438 male, 192 female), 6300 utterances, 8 dialect regions	Language: English,Type: Artificial, Modality: Aud, Txt	Gender, Age	1987
OGI [54]	Total 1927 telephone calls, Average 175 calls per language	Language: Multilingual (11), Type: Natural, Modality: Aud	Gender	1994
IViE Corpus [62]	110 speakers (54 male, 56 female), 9 dialect regions	Language: British English, Type: Artificial, Modality: Aud, Txt	Gender	2001
Dutch Corpus [64]	308.3 recorded hours, 555 speakers (436 male, 19 female), age groups(18–81)	Language: Dutch(2), Type: Artificial, Natural, Modality: Aud	Gender, Age	2008
NIST SRE08 [230]	942 recorded hours, telephone and microphone data	Language: Multilingual(25), Type: Artificial, Natural, Modality: Aud	Age	2011
Common Voice [231]	30,329 recorded hours, age groups (20–79), 45% male, 17% female, 2% others	Language: Multilingual (120), Type: Artificial, Natural, Modality: Aud, Txt	Gender, Age, Accent	2019
SUSAS [232]	4 Emotions × 32 Speakers(19 male, 13 female), 16 000 utterances	Language: English, Type: Artificial, Natural, Modality: Aud	Emotion	1999
EMO-DB [97]	7 Emotions × 10 Speakers(5 male, 5 female), 10 utterances	Language: German, Type: Artificial, Modality: Aud	Emotion	2005
CASIA [103]	6 Emotions × 4 Speakers(2 male, 2 female), 9600 utterances	Language: Mandarin, Type: Artificial, Modality: Aud	Emotion	2008
IEMOCAP [120]	9 Emotions × 10 Speakers(5 male, 5 female), 10039 utterances	Language: English, Type: Artificial, Modality: Aud	Emotion	2008
TESS [116]	8 Emotions × 2 Speakers(female), 2800 utterances	Language: English, Type: Artificial, Modality: Aud	Emotion	2010
RECOLA [233]	2 Emotions × 46 Speakers(27 male, 13 female), 9.5 h	Language: French, Type: Artificial, Modality: Aud, Vis, ECG, EDA	Emotion	2013
CREMA-D [234]	6 Emotions × 91 Speakers(48 male, 43 female), 7442 utterances	Language: English, Type: Artificial, Modality: Aud, Vis, Aud-Vis	Emotion	2014
SAVEE [117]	8 Emotions × 4 Speakers(male), 480 utterances	Language: English, Type: Artificial, Modality: Aud-Vis	Emotion	2015
CHEAVD [235]	6 Emotions × 238 Speakers, 140 min emotional segments	Language: Chinese, Type: Artificial, Modality: Aud-Vis	Emotion	2016
RAVDESS [236]	7 Emotions × 24 Speakers(12 male, 12 female), 7356 utterances	Language: Multilingual (2), Type: Artificial, Modality: Aud, Vis, Aud-Vis	Gender, Emotion	2018
DAIC-WOZ [146]	50.4 recorded hours, 189 interviewed healthy + controls	Conditions: anxiety, depression, post-traumatic stress disorder, Modality: Aud, Vis, Txt	Mental Health	2014
MODMA [144]	23 clinically depressed patients + 29 healthy controls, Demographic data, Psychological assessments	Conditions: mental disorders, Modality: Aud, EEG	Mental Health	2015
SVD [237]	2000 speakers(687 healthy, 1356 pathological cases), Vowel Recordings, Pitch recordings, Sentence recordings	Conditions: 71 different pathologies, Modality: Aud, EGG	Voice Pathology	2012
MEEI [238]	139 speakers(53 healthy, 86 pathological cases), Vowel Recordings, Running speech recordings, Isolated words recordings	Language: English, Conditions: 6 vocal disorders, Modality: Aud	Voice Pathology	2017
AVPD [239]	50 speakers(25 healthy, 25 pathological cases), Vowel Recordings, Running speech recordings, Isolated words recordings	Language: Arabic, Conditions: 4 vocal disorders, Modality: Aud	Voice Pathology	2017
AMI Corpus [240]	100 recorded hours, Monologues, Dialogs and multi-party discussions recording	Language: English, Personality measure: Big Five, Modality: Aud, Vis	Personality	2005
First Impression [241]	41.6 recorded hours, 10,000 labeled video clips	Language: English, Personality measure: Big Five, Modality: Vis	Personality	2017

(*continued on next page*)

(FSD50K) [242], offer extensive coverage of acoustic scenes and event detection tasks. AudioSet, with over 2 million human-labeled audio clips spanning 632 classes, is unparalleled in scale and diversity, making it a go-to resource for broad-spectrum audio analysis. However, both datasets face challenges such as class imbalance, which can hinder model performance on underrepresented categories.

A growing trend in audio research is the adoption of multimodal datasets that integrate audio with other modalities. Datasets such as

Table 7 (*continued*).

Dataset	Composition	Characteristics	AP tasks	Year
FSD50K [242]	51,197 audio clips, sounds produced by physical sound sources and production mechanisms	Number of classes: 200, Total length: 108.3 h, Class proportions: Imbalanced, Modality: Aud	ASC, AED	2022
DESED [243]	10 s sound clips, sounds recorded or synthesized to simulate a domestic environment	Number of classes: 10, Total length: >12 h, Class proportions: Imbalanced, Modality: Aud	AED	2019
TUT Acoustic Scenes [244]	10 s sound clips, sounds of indoor, outdoor environments	Number of classes: 10, Total length: 24 h, Class proportions: Balanced, Modality: Aud	ASC	2018
Audio Set [245]	10 s sound clips, 2.1 million annotated Youtube videos	Number of classes: 527, Total length: 4971 h, Class proportions: Maximally balanced, Modality: Aud, Vis	ASC, AED	2017
MAD [246]	1–10 s sound clips, 8075 samples, high levels of background noise, realistic military scenarios	Number of classes: 7, Total length: 12 h, Class proportions: Maximally balanced, Modality: Aud	AED	2024
ESC-50 [227]	5 s sound clips, sounds of animals, natural soundscapes, non-speech sounds, domestic and urban sounds	Number of classes: 50, Total length: 2.7 h, Class proportions: Balanced, Modality: Aud	ASC	2015
UrbanSounds8K [247]	8732 labeled sound excerpts (≤4 s) of urban sounds	Number of classes: 10, Total length: 27 h, Class proportions: Maximally Balanced, Modality: Aud	ASC, AED	2014

Fig. 7. Use cases for audio profiling across various domains.

DAIC-WOZ [146], MODMA [144], and SVD [237] combine audio with visual data, electroencephalography (EEG), or electroglottography (EGG). These datasets are increasingly used for mental health diagnostics, providing a richer, multidimensional perspective on human behavior. For instance, DAIC-WOZ has been instrumental in depression detection research due to its detailed interview recordings and associated clinical scores. Selecting an appropriate dataset requires careful consideration of factors such as demographic diversity, class balance, and task-specific requirements.

7. Applications and use cases

Audio profiling has a lot of practical use case scenarios. Having the ability to recognize various physical traits such as age, gender, height, weight, face, and physiological traits such as emotional and mental state, and social traits such as socio-economic status [248–250], religion, ethnicity, and dialect can be invaluable in various fields. Based on several works on user profiling through voice, some of the popular areas where profiling can play significant roles are shown in Fig. 7.

7.1. Law and security

Profiling can be used as a supportive tool for the investigation of crimes where voice is one of the primary evidence. Crimes such as blackmailing, voice phishing (vishing), extortion, hoaxes and pranks, harassment, and threats are committed using voice. With technology-assisted crimes on the rise, especially those involving phone calls, profiling can predict key traits of the perpetrator, potentially aiding in quicker identification than otherwise possible.

Voice signals exhibit a vast number of unique parameters, making no two voices alike worldwide. This distinctive nature makes them well suited for biometric security applications. Currently, voice biometrics are employed in speaker identification, where a speaker is recognized from a set of possible candidates, and speaker verification or authentication, which involves confirming a speaker's claimed identity. These tasks rely on creating a voice signature and matching it against stored signatures in a database. However, studies have shown that speaker-matching applications can be vulnerable to adversarial attacks designed to spoof voice authentication [251–253]. To strengthen security, integrating audio profiling can provide an additional layer of protection by analyzing attributes such as age, accent, or gender to verify if they align with the registered identity. A practical application of voice biometrics can be seen in ANZ Bank's Voice ID system [254]. This technology analyzes a caller's voice and matches it against a database of registered clients. If a match is found, the system authenticates the user, granting access to banking services. ANZ's Voice ID enables customers to make secure payments over $1000 via their mobile app without requiring additional passwords or PINs, leveraging voice biometric technology for enhanced security and convenience. Beyond identity verification, incorporating audio profiling could further improve fraud detection by assessing variations in speech patterns or contextual cues, making authentication systems more resilient to spoofing attempts.

With the support of ASC/AED systems, deeper insights can be gained into the environment of interest. For example, it is possible to determine if a perpetrator is calling from inside a car or public transportation, such as a train or tram, which could aid in fugitive cases. Furthermore, ASC/AED systems can be employed in surveillance, helping identify potential threats or events like glass breaking, gunshots, or abnormal sounds [222]. In certain restricted environments, such as prisons, where video surveillance is legally limited, audio monitoring can serve as a viable alternative for detecting illegal activities within legal boundaries [255].

Facial reconstruction from voice is an active research field focused on generating facial features or entire faces from voice signals. This area holds significant potential in forensics, as reconstructing facial features from voice—and vice versa—can provide valuable insights. Voice is both directly and indirectly linked to physical characteristics, such as bone structure, height, weight, age, and gender. These relationships enable predictive modeling, where parameters like body mass index and skeletal proportions can be inferred from voice analysis. Leveraging this knowledge, it is technologically feasible to create a 3D reconstruction of the face and even the entire body. Detailed discussions on facial reconstruction from voice can be found in the article by [256].

7.2. HealthCare

Profiling can be applied in healthcare for non-invasive assessment and monitoring of physical and mental health. It is useful for tracking medical drug compliance, detecting intoxication, substance abuse, and various health indicators. Telemonitoring, especially for underprivileged populations, can provide early warnings of diseases and improve treatment outcomes. It can be used for faster diagnosis of different voice-related pathologies [160,161]. Voice contains a rich array of information and can be a potential biomarker for various diseases. A real-world application of audio profiling in healthcare is **Sleep.ai**, a system developed by AltexSoft for a Dutch healthcare startup. This solution detects teeth grinding and snoring sounds during sleep to help dentists identify and monitor bruxism—a condition where patients unconsciously clench or grind their teeth. By tracking sleep-related sounds, Sleep.ai enables early detection and long-term monitoring, allowing healthcare professionals to understand the root causes of bruxism and recommend appropriate treatment strategies [257].

Another example of audio profiling application is Hume AI's empathic voice interface (EVI 2), which enables large language models (LLMs) to detect and express human emotions. This technology analyzes vocal cues to infer emotions such as anxiety, determination, or happiness and responds with contextually appropriate emotional tones. For instance, if a user speaks in a sad tone, the system can recognize this and provide a sympathetic response, making AI interactions more natural and emotionally engaging. This approach enhances user engagement in applications such as virtual assistants, customer service, and mental health support. For further details, see the article on Wired. An interesting recent use case involves leveraging audio profiling to support call center workers by transforming angry customer tones into calmer, more neutral ones during live interactions. The system is designed to detect vocal stress markers such as pitch and inflection associated with anger, adjusting them in real-time while preserving the essential emotional context to ensure appropriate responses. This technology aims to alleviate the psychological stress caused by "kasuhara" (customer harassment), a significant and growing issue in Japan. SoftBank Corp., a leading telecommunications provider in Japan, plans to commercialize the system by 2026, with the goal of enhancing worker well-being and productivity in high-stress environments. For more details, see the article on Reuters. The Section 5.1.5 gives a distilled knowledge about voice pathology detection. Additionally, Section 5.1.4 discusses several mental health issues that can be addressed using voice.

7.3. Commercial and social services

Profiling technology has the potential to revolutionize interactions in business and educational settings. It can monitor and flag calls for issues such as anger, dissatisfaction, fraud, and intoxication, helping to prevent business losses. Additionally, it can enhance persuasion and leadership skills across various professions and monitor professionals for signs of fatigue.

A patent by Amazon Technologies Inc. [24] describes a system that extracts real-time physical and emotional traits from user voice inputs to identify specific conditions or characteristics. This enables tailored content suggestions or service offerings aligned with the user's immediate state. For instance, the system can detect a cough and suggest medications accordingly. However, health-related data, classified as sensitive under Article 9 of the European Union's General Data Protection Regulation (GDPR), requires strict protections due to its sensitivity. Prejudiced applications, such as adjusting insurance premiums based on inferred health status, remain a concern [258].

A facet of profiling technology is the automated personality recognition systems, which have a wide array of applications across industries. These systems enhance personal assistants like Siri and Alexa by enabling tailored responses based on user personality. They also improve recommendation systems, assist in sentiment analysis and word polarity detection, provide personalized counseling in healthcare, and support forensic investigations and deception detection. Additionally, they optimize job screening processes, advance psychological studies by analyzing behavior-personality connections, and inform political campaigns by creating targeted voter profiles for more effective persuasion strategies [185]. For instance, audio profiling technology is increasingly used in recruitment. By analyzing voice characteristics, organizations can identify candidates best suited for specific roles. Some companies assess candidates' reliability and positivity through voice analysis, helping select individuals with the right mindset for the job [258].

Beyond recruitment, ASC/AED enable smart devices to recognize and respond to specific sounds. For example, a smart home system can turn on lights upon detecting a doorbell or identify alarms and emergencies [222]. Overall, audio profiling holds significant potential to enhance organizational efficiency and drive revenue growth across industries.

7.4. Entertainment and games

Profiling technology has exciting applications in entertainment and recreation, such as creating 2D and 3D representations of individuals based on their voice. It can enable customized gaming experiences tailored to players' traits and preferences and create themed characters in entertainment and amusement park settings. Additionally, acoustic event detection can enhance virtual reality experiences by providing more immersive audio environments that react dynamically to different sound events.

8. Open issues and future directions

In this section, we discuss the prominent open issues that hinder the effectiveness of Audio Profiling. We discuss issues with the attribute inference process, dataset availability and **privacy preservation**. It also explores future research directions to address these challenges. Identifying and analyzing these challenges is of utmost importance when seeking novel and technical solutions.

8.1. Issues in attributes inferences

The literature on personal attribute inference, mainly in terms of gender and age prediction, has several issues, such as efficiency discrepancies between genders [51,61,70] indicating potential biases or challenges in accurately capturing vocal characteristics. A consensus is lacking regarding the specific features or feature sets that prominently characterize each attribute. Additionally, the challenge of accurately predicting age, emotion, gender, height, etc. for voices with various accents, speech impairments, or other unique characteristics emphasizes the potential complexity of generalizing models to diverse linguistic characteristics. Furthermore, the inability to detect attributes accurately when audio clips consist of voices from more than one person simultaneously highlights the challenges in handling multispeaker scenarios. There exists some difference between the perceived age and actual age, and also voice patterns are affected by multiple parameters, such as weight, height, and emotions [63].

GMMs are widely used in age classification. The major issues are primarily related to the varying effectiveness of different speech features and the number of mixtures used in the GMM method, as well as the challenges in accurately classifying adult voices compared with children's voices.[82]. Different end-to-end learning models that use voice embeddings such as x-vectors and i-vectors have been used for age prediction and classification. Although effective for the task at hand, these approaches face challenges related to **data dependency**: they require a large amount of training data, which might not always be available. In addition, the study [41] suggests that the DNN back-end might struggle with high-dimensional inputs, as seen in the limited improvement when using i-vectors alone for the back-end, but improved performance when concatenated with x-vectors. This limitation could affect the scalability and adaptability of the system. Many studies rely heavily on **Transfer Learning**, which might not capture the full diversity of human voices, especially in different linguistic or cultural contexts [70].

These difficulties underscore the necessity for a thorough understanding and refinement of models to guarantee an unbiased and precise inference of attributes from voice data. Future direction lies in enhancing **multi-attribute inference models.** Efforts to prioritize the development of unbiased systems to mitigate efficiency discrepancies in multi-attribute inference, ensuring equitable performance across diverse demographics is essential. Addressing the complexities of multi-speaker scenarios and improving model generalization for voices with various accents, speech impairments, and cultural variations are essential. Additionally, refining transfer learning techniques to reduce dependency on large datasets and exploring robust feature sets, such as d-vector embeddings [85] can significantly enhance model accuracy and scalability.

8.2. Quality and type of data

8.2.1. Dataset availability

The efficacy of Audio profiling relies completely on high quality of datasets available. Audio profiling is about personal attributes and environmental context detection that need high quality and quantity data. In the case of personal traits detection, currently available dataset can be classified as: **Natural datasets, Semi-Natural and Actor Based Dataset**. Natural datasets, such as recordings from reality shows, youtube vlogs, customer service calls contain pure and unaltered characteristics but are often restricted due to ethical and privacy concerns. The next type of dataset is the semi-natural dataset that contains audio clips/recordings from scenarios that are real and the speakers have no idea of being recorded. Datasets like [232] is an example. The most common dataset is the actor-based dataset which contains recordings from professional actors who imitate different sounds as needed. The prominent issue with datasets is the absence of full variability representation encountered in the real world. The authenticity and practicality of such a dataset need to be properly scrutinized by future research.

Datasets for other attributes such as weight estimation, body build parameters like shoulder width, waist size, etc., along with voice recordings are difficult to find. Additionally, these estimations may not be reliable due to the less direct relationship between these parameters and the speech signal. The limited availability of datasets that cover a wide range of linguistic, cultural, and demographic diversity restricts the generalizability of models.

8.2.2. Data imbalance

Many datasets show disparities in the gender and age representation of speakers. For instance, the TIMIT dataset has 192 female and 438 male speakers, Dutch Corpus [64] has uneven 436 male speakers and only 19 female speakers. While the VoxCeleb dataset [259], though more balanced, still has a skew with 55% male and 45% female speakers. These imbalances can significantly impact the performance and generalizability of audio profiling systems, especially when considering variations in languages, accents, and emotional expressions.

To address these challenges, future research should prioritize the creation of diverse, high-quality audio datasets. Additionally, efforts should focus on developing **universal benchmark datasets** that accurately capture real-world variability across linguistic and paralinguistic dimensions. One promising avenue for addressing dataset limitations is extending advanced audio synthesis frameworks, such as AudioLM [260]. These models offer the capability to generate synthetic audio that closely mimics the nuances of human speech and environmental sounds, thereby overcoming the constraints of limited real-world data. Future research should explore the development of domain-specific adaptations of Large Audio Models **(LAMs)** to produce datasets tailored for specialized tasks such as emotion recognition, pathological speech analysis, or environmental sound classification.

8.3. Model complexity and computational requirements

Complex Architectures those incorporating deep learning such as LSTM-RNNs, DNNs, Transformers are computationally intensive and complex, requiring substantial resources for training and deployment. Tuning and optimizing these complex model, including tuning hyperparameters and network layers, for specific tasks can be challenging and resource-consuming. Audio Profiling in general requires heavy computational resources. As the current models are getting complex and the parameters increasing exponentially, limitations of computational resource is another crucial issue. There is a need for developing models that are more resource-efficient, making audio profiling accessible to systems with limited computational capacity. Future research should aim to **streamline model architectures** to make them more resource-efficient. Current state-of-the-art architectures, such as transformers, often require substantial computational resources. Future research should explore lightweight architectures capable of delivering high accuracy while reducing computational and energy demands. Possible exploration of architecture like **xLSTM** [261], which have proven efficient in low-resource settings, merits investigation for audio profiling. These advancements would facilitate the deployment of audio profiling systems in resource-constrained environments and make the technology more accessible.

8.4. Issues related to real-world application and integration with real-time systems

Many systems are tested in controlled conditions, and their adaptability to real-world, noisy, and dynamic environments is often untested. Systems may not generalize well across different speech contexts, languages, or accents. Furthermore, integrating these models into real-time systems (like call centers or interactive voice response systems) poses challenges in terms of latency and continuous streaming of data. Many models are designed for specific utterance lengths, and their effectiveness with varying speech duration can be limited.

8.5. Issues beyond computational complexity

Human voice is complex and not deterministic. Even when conditions are constant, the same person can produce different voice sounds. This variability is affected by emotions, mental and physical states, social and environmental circumstances, making each person unique. Thus, **Inherent Human Variability** is both a challenge and opportunity for profiling.

Human poses another set of skill, intentional alteration of their voice. **Voice Disguise** includes impersonation and masking. Impersonation and mimicry involve humans altering their voices to resemble someone else, though the former generally carries a more negative connotation compared to the latter. Identifying an impersonated voice is an issue as the judgments about the speaker are likely to be incorrect. The challenge for accurate profiling is to find elements in a voice that are not under voluntary control of the speaker, and do not change with disguise, and derive conclusions based on those aspects. For this to happen, a system to identify disguised voice is needed. Thus, Audio profiling systems that are fool proof from voice disguise remains an open problem. The author of the book [1] suggests that employing **micro-articulometry** represents a logical approach to addressing the shortcomings of profiling when dealing with voice disguise.

Voice Masking refers to the concealment of the natural attributes of one's voice, suppressing its genuine qualities. This can involve imitating sounds like that of animals, musical instruments, or even external aids such as physical masks or electronic voice transformation devices. As long as the fundamental basis remains natural human speech, the potential for successful profiling exists, although the efficacy varies for different types of masking. On the contrary, if voice or speech is entirely synthesized, profiling is likely to, and ideally should, fail. Currently, achieving an accurate synthesis of human voice or speech, including its micro-nuances, without borrowing elements from real voices is challenging [1]. With the recent AI breakthroughs, voice synthesizing tools are readily available [262] which necessitates systems capable of detecting and handling synthetic voices. Detecting synthetic voices in isolation, similar to addressing the issue of voice disguise or masking, is a concern that must be tackled within the context of profiling. Future research should investigate methods to identify and counteract these alterations for generalized audio profiling models. Emotion-resilient models that differentiate between intrinsic voice attributes and transient emotional states could ensure consistent profiling under diverse conditions.

8.6. Privacy concerns and preservation

Privacy preservation refers to the measures, techniques, and strategies employed to protect sensitive data from unauthorized access, disclosure, and misuse. It aims to ensure that personal and confidential data remains secure and is only accessible to authorized entities. Privacy preservation is vital in many fields, including healthcare, finance, and communication. For example, in today's digital era, with voice assistants like Alexa, Google Assistant, Siri, and Cortana integrated into devices such as smart speakers, smartphones, and IoT systems, protecting privacy in voice user interfaces is essential. Voice assistants do not only process spoken commands but they can also extract paralinguistic information such as emotions, gender, and even potential health conditions [24]. An article published by Yale University Press examines the emergence of audio profiling and its potential impacts on personal freedom. The study highlights privacy concerns over how companies collect and analyze voice data to infer personal traits. For instance, without proper privacy measures, an insurance company might analyze a client's voice, detect a health issue, and raise client premiums without his/her consent. This is why strong privacy preservation is needed to keep all aspects of voice data secure.

In the literature, various approaches have been proposed to address privacy concerns and protect paralinguistic information embedded in speech data such as emotional states, gender identity, and personal attributes. One key strategy for voice privacy preservation is limiting transmitted information to what is essential for downstream tasks—a concept known as **information isolation**, where channels of information separate target information from private or undesired data. The **Information Bottleneck** (IB) [263] method achieves information isolation by compressing speech data through an encoder, creating a constrained bottleneck that transmits only essential information. A decoder then reconstructs data for the task. Techniques like dimensionality reduction and quantization ensure private information is excluded. The IB principle is effective in applications such as anonymizing speech attributes [264] and learning low-bitrate representations that retain semantic content [263]. The strengths of the IB method include its provable privacy guarantees (determined by the bottleneck's bitrate), flexibility in balancing privacy and utility, and adaptability to various tasks. However, challenges include designing appropriate loss functions for tight bottlenecks, vulnerability to adversarial attacks, and ensuring robust system design to maintain task accuracy while discarding private data [265].

Another approach used by researchers to protect data privacy is **Adversarial Learning** [266,267]. In this approach, models are trained to balance utility and privacy through competing objectives. A trusted task (e.g., speaker verification) extracts task-specific features, while an adversarial task (e.g. gender classification) penalizes the system for retaining sensitive attributes. This adversarial competition forces the encoder to suppress private information, such as emotion or gender, while preserving utility. For instance, CycleGAN has been used to neutralize emotional attributes in speech [268], while GenGAN synthesizes gender-ambiguous voices to enhance privacy [269]. GenGAN employs a U-Net generator and an AlexNet-based discriminator, leveraging adversarial loss to balance utility and privacy with minimal distortion. Validation on LibriSpeech [270] shows it outperforms methods like VQ-VAE-Gen [271] and PCMelGAN [272] in word recognition accuracy while reducing gender and identity inference rates.

Disentanglement Learning is another approach for data privacy preservation that leverages disentangled representations to separate observed data into distinct, independent features, enhancing robustness, interpretability, and generalization. In computer vision, this approach has been applied to tasks like pose-invariant recognition, adversarial disentanglement for attribute transfer [273], and person re-identification [274]. Beyond vision, disentangled representations help mitigate bias for fairness and facilitate domain adaptation by separating domain-specific features [275]. Many research studies [271,276–278] have used speech disentanglement to separate signals into independent channels—such as linguistic content, speaker identity, and emotions—enabling selective sharing of private information tailored to application requirements.

Other approaches include voice transformation for privacy, such as **Voice Sanitization** [279], speech anonymization using the **McAdams coefficient** [280], and the **VoicePM** framework [281], which provides a structured methodology for evaluating privacy-utility trade-offs across different anonymization techniques. VoicePM systematically assesses anonymization methods such as signal processing, voice conversion, and adversarial perturbation, measuring their effectiveness in concealing speaker attributes while preserving intelligibility and downstream task performance.

Although these methods demonstrate significant progress in safeguarding sensitive voice attributes, they also highlight persistent challenges in paralinguistic privacy preservation. Key issues include: (1) achieving a balance between privacy protection and data usability, as overly strict measures risk compromising the practical utility of audio [282], (2) incomplete anonymization remains a persistent issue, with residual sensitive information potentially enabling re-identification [277,283], and (3) the lack of standardized evaluation metrics for assessing privacy-utility trade-offs, which hinders consistent progress and comparison across approaches [281,282].

Researchers must develop more sophisticated, efficient, and adaptable methods that effectively address these challenges and ensure robust protection of sensitive speech and audio data. A promising direction in voice anonymization is the integration of Neural Audio Codecs (NACs), as introduced by [284]. NACs, such as EnCodec [285], represent speech as discrete acoustic tokens, which can be manipulated to anonymize speaker identity while preserving semantic content. Unlike conventional methods that rely on perturbing speaker embeddings, NACs offer a more consistent and effective approach to speaker identity concealment [284]. NACs employ quantized codes to bottleneck speaker-related information, ensuring a more robust and reliable anonymization process. Additionally, NACs can also be integrated with audio language models such as VALL-E [286] and AudioLM [260] to generate high-quality synthetic speech that maintains linguistic integrity while effectively concealing speaker identity. These models leverage discrete acoustic representations, enabling robust anonymization of speech attributes without introducing significant distortions to linguistic content. Future research should focus on improving these methods to strengthen privacy guarantees.

9. Conclusion

In this survey, we have provided an up-to-date review of the audio profiling paradigm. We have described the audio profiling pipeline and highlighted the datasets, features, methods, and various audio profiling tasks. We have also discussed several attributes that can be inferred from voice recordings and how these combine to paint the whole picture of a speaker in the recording. This study has compiled and focused on the diverse and significant applications of audio profiling systems. In addition to the comprehensive review, we have highlighted the challenges that impede the effectiveness of profiling systems, with an emphasis on privacy preservation, thereby providing directions for future research in this field. We aspire for this article to serve as a definitive resource for researchers and practitioners engaging with audio profiling. As voice recording devices become increasingly embedded in the fabric of a smart society and advancements in audio profiling systems continue to rise, these technologies must be safeguarded. The future of audio profiling lies in addressing these multifaceted challenges through innovative research, interdisciplinary collaboration, and a focus on ethical, practical, and scalable solutions. The goal is to develop systems that are unbiased, efficient, and respectful of user privacy, thereby enhancing the reliability and applicability of audio profiling technologies in diverse real-world scenarios.

CRediT authorship contribution statement

Anil Pudasaini: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Muna Al-Hawawreh:** Writing – review & editing, Validation, Supervision. **Mohamed Reda Bouadjenek:** Supervision, Project administration. **Hakim Hacid:** Funding acquisition, Conceptualization. **Sunil Aryal:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is supported by the Technology Innovation Institute, UAE under the research contract number TII/DSRC/2022/3143.

Data availability

No data was used for the research described in the article.

References

[1] R. Singh, Profiling Humans from their Voice, Springer Singapore, Singapore, 2019, http://dx.doi.org/10.1007/978-981-13-8403-5, URL http://link.springer.com/10.1007/978-981-13-8403-5.

[2] J.L. Kröger, O.H.-M. Lutz, P. Raschke, Privacy implications of voice and speech analysis – information disclosure by inference, in: M. Friedewald, M. Önen, E. Lievens, S. Krenn, S. Fricker (Eds.), Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers, in: IFIP Advances in Information and Communication Technology, Springer International Publishing, Cham, 2020, pp. 242–258, http://dx.doi.org/10.1007/978-3-030-42504-3_16.

[3] A. Bastanfard, D. Amirkhani, M. Hasani, Increasing the accuracy of automatic speaker age estimation by using multiple UBMs, in: 2019 5th Conference on Knowledge Based Engineering and Innovation, KBEI, 2019, pp. 592–598, http://dx.doi.org/10.1109/KBEI.2019.8735005.

[4] L. Jasuja, A. Rasool, G. Hajela, Voice gender recognizer recognition of gender from voice using deep neural networks, in: 2020 International Conference on Smart Electronics and Communication, ICOSEC, 2020, pp. 319–324, http://dx.doi.org/10.1109/ICOSEC49089.2020.9215254.

[5] X. Chen, Z. Pan, A convenient and low-cost model of depression screening and early warning based on voice data using for public mental health, Int. J. Environ. Res. Public Heal. 18 (12) (2021) http://dx.doi.org/10.3390/ijerph18126441, URL https://www.mdpi.com/1660-4601/18/12/6441.

[6] N. Cummins, A. Baird, B.W. Schuller, Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning, Methods 151 (2018) 41–54, http://dx.doi.org/10.1016/j.ymeth.2018.07.007, URL https://www.sciencedirect.com/science/article/pii/S1046202317303717.

[7] R.A. Khalil, E. Jones, M.I. Babar, T. Jan, M.H. Zafar, T. Alhussain, Speech emotion recognition using deep learning techniques: A review, IEEE Access 7 (2019) 117327–117345, URL https://api.semanticscholar.org/CorpusID:201833712.

[8] S. Hegde, S. Shetty, S. Rai, T. Dodderi, A survey on machine learning approaches for automatic detection of voice disorders, J. Voice 33 (6) (2019) 947.e11–947.e33, http://dx.doi.org/10.1016/j.jvoice.2018.07.014, URL https://www.sciencedirect.com/science/article/pii/S0892199718301437.

[9] X. Xia, R. Togneri, F. Sohel, Y. Zhao, D. Huang, A survey: Neural network-based deep learning for acoustic event detection, Circuits Systems Signal Process. 38 (2019) 3433–3453.

[10] J. Abeßer, A review of deep learning based methods for acoustic scene classification, Appl. Sci. 10 (6) (2020) http://dx.doi.org/10.3390/app10062020, URL https://www.mdpi.com/2076-3417/10/6/2020.

[11] M.S. Fahad, A. Ranjan, J. Yadav, A. Deepak, A survey of speech emotion recognition in natural environment, Digit. Signal Process. 110 (2021) 102951, http://dx.doi.org/10.1016/j.dsp.2020.102951, URL https://www.sciencedirect.com/science/article/pii/S1051200420302967.

[12] J. Shen, J. Cao, O. Lederman, S. Tang, A.S. Pentland, User profiling based on nonlinguistic audio data, ACM Trans. Inf. Syst. 40 (1) (2021) http://dx.doi.org/10.1145/3474826.

[13] N. Abdulmajeed, B. Al-Khateeb, M. Mohammed, A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions, J. Intell. Syst. 31 (1) (2022) 855–875, http://dx.doi.org/10.1515/jisys-2022-0058.

[14] U.H. Jaid, A.K.A. Hassan, Review of automatic speaker profiling: Features, methods, and challenges, Iraqi J. Sci. (2023) 6548–6571, http://dx.doi.org/10.24996/ijs.2023.64.12.36, URL https://ijs.uobaghdad.edu.iq/index.php/eijs/article/view/7972.

[15] A. Hashem, M. Arif, M. Alghamdi, Speech emotion recognition approaches: A systematic review, Speech Commun. 154 (2023) 102974, http://dx.doi.org/10.1016/j.specom.2023.102974, URL https://www.sciencedirect.com/science/article/pii/S0167639323001085.

[16] L. Yue, P. Hu, J. Zhu, Advanced differential evolution for gender-aware English speech emotion recognition, Sci. Rep. 14 (1) (2024) 17696, http://dx.doi.org/10.1038/s41598-024-68864-z.

[17] I. Shahin, Gender-dependent emotion recognition based on HMMs and SPHMMs, 2018, http://dx.doi.org/10.1007/s10772-012-9170-4, arXiv:1801.06657.

[18] N.P. Trilok, S.-H. Cha, C.C. Tappert, Establishing the uniqueness of the human voice for security applications, in: Proceedings CSIS Research Day, Pace University, NY, May, 2004.

[19] S.M. Hughes, M.J. Pastizzo, G.G. Gallup, The sound of symmetry revisited: Subjective and objective analyses of voice, J. Nonverbal Behav. 32 (2008) 93–108, http://dx.doi.org/10.1007/s10919-007-0042-6.

[20] S.M. Hughes, F. Dispenza, G.G. Gallup, Ratings of voice attractiveness predict sexual behavior and body configuration, Evol. Hum. Behav. 25 (5) (2004) 295–304, http://dx.doi.org/10.1016/j.evolhumbehav.2004.06.001, URL https://www.sciencedirect.com/science/article/pii/S109051380400042X.

[21] M. Zuckerman, R.E. Driver, What sounds beautiful is good: The vocal attractiveness stereotype, J. Nonverbal Behav. 13 (1988) 67–82, http://dx.doi.org/10.1007/bf00990791.

[22] S.M. Hughes, B.C. Rhodes, Making age assessments based on voice: The impact of the reproductive viability of the speaker, J. Soc. Evol. Cult. Psychol. 4 (2010) 290–304, http://dx.doi.org/10.1037/h0099282.

[23] Y. Wu, T. Lee, Time-frequency feature decomposition based on sound duration for acoustic scene classification, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2020, pp. 716–720, http://dx.doi.org/10.1109/ICASSP40776.2020.9053194.

[24] H. jin, S. Wang, Voice-based determination of physical and emotional characteristics of users, 2017, URL https://patentimages.storage.googleapis.com/f6/a2/36/d99e36720ad953/US10096319.pdf.

[25] T. Heittola, E. Çakır, T. Virtanen, The machine learning approach for analysis of sound scenes and events, in: Computational Analysis of Sound Scenes and Events, Springer International Publishing, Cham, 2018, pp. 13–40, http://dx.doi.org/10.1007/978-3-319-63450-0_2.

[26] J. Schröder, N. Moritz, M.R. Schädler, B. Cauchi, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, S. Goetze, On the use of spectro-temporal features for the IEEE AASP challenge 'detection and classification of acoustic scenes and events', in: 2013 IEEE Workshop on Applications of Signal Processing To Audio and Acoustics, IEEE, 2013, pp. 1–4.

[27] T. Heittola, A. Mesaros, T. Virtanen, M. Gabbouj, Supervised model training for overlapping sound events based on unsupervised source separation, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 8677–8681.

[28] B. Zellner, Pauses and the temporal structure of speech, in: Zellner, B. (1994). Pauses and the Temporal Structure of Speech, in E. Keller (Ed.) Fundamentals of Speech Synthesis and Speech Recognition.(Pp. 41-62). Chichester: John Wiley, John Wiley, 1994, pp. 41–62.

[29] F. Alías, J.C. Socoró, X. Sevillano, A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds, Appl. Sci. 6 (5) (2016) http://dx.doi.org/10.3390/app6050143, URL https://www.mdpi.com/2076-3417/6/5/143.

[30] K. Becker, Gender Recognition by Voice, URL https://www.kaggle.com/datasets/primaryobjects/voicegender.

[31] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. Audio Speech Lang. Process. 19 (4) (2011) 788–798, http://dx.doi.org/10.1109/TASL.2010.2064307, URL https://ieeexplore.ieee.org/abstract/document/5545402. Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.

[32] N. Dehak, S. Shum, Low-dimensional speech representation based on Factor Analysis and its applications!, 2011, URL https://people.csail.mit.edu/sshum/talks/ivector_tutorial_interspeech_27Aug2011.pdf.

[33] Y.-A. Chung, J. Glass, Speech2Vec: A sequence-to-sequence framework for learning word embeddings from speech, 2018, arXiv:1803.08976. URL https://arxiv.org/abs/1803.08976.

[34] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-Vectors: Robust DNN embeddings for speaker recognition, IEEE Press (2018) 5329–5333, http://dx.doi.org/10.1109/ICASSP.2018.8461375.

[35] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, B. Schuller, Snore sound classification using image-based deep spectrum features, 2017.

[36] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, S. Khudanpur, Speaker recognition for multi-speaker conversations using X-vectors, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019, pp. 5796–5800, URL https://api.semanticscholar.org/CorpusID:146117456.

[37] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L.P. García-Perera, F. Richardson, R. Dehak, P.A. Torres-Carrasquillo, N. Dehak, State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations, Comput. Speech & Lang. 60 (2020) 101026, http://dx.doi.org/10.1016/j.csl.2019.101026, URL https://www.sciencedirect.com/science/article/pii/S0885230819302700.

[38] S.S. Stevens, J. Volkmann, The relation of pitch to frequency: A revised scale, Am. J. Psychol. 53 (3) (1940) 329–353.

[39] A.A. Alnuaim, M. Zakariah, C. Shashidhar, W.A. Hatamleh, H. Tarazi, P.K. Shukla, R. Ratna, Speaker gender recognition based on deep neural networks and ResNet50, in: M.F. Hashmi (Ed.), Wirel. Commun. Mob. Comput. 2022 (2022) 1–13, http://dx.doi.org/10.1155/2022/4444388.

[40] A.M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M.Y. Lee, S. Kwon, S.W. Baik, Deep features-based speech emotion recognition for smart affective services, Multimedia Tools Appl. 78 (5) (2019) 5571–5589, http://dx.doi.org/10.1007/s11042-017-5292-7.

[41] P. Ghahremani, P.S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, N. Dehak, End-to-end deep neural network age estimation, in: Interspeech, vol. 2018, 2018, pp. 277–281.

[42] M. Kaushik, V.T. Pham, E.S. Chng, End-to-end speaker height and age estimation using attention mechanism with LSTM-RNN, 2021, arXiv:2101.05056. URL https://arxiv.org/abs/2101.05056.

[43] G.S. Liu, J.M. Hodges, J. Yu, C.K. Sung, E. Erickson-DiRenzo, P.C. Doyle, End-to-end deep learning classification of vocal pathology using stacked vowels, Laryngoscope Investig. Otolaryngol. 8 (5) (2023) 1312–1318, http://dx.doi.org/10.1002/lio2.1144, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/lio2.1144. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lio2.1144.

[44] S.K. Pandey, H.S. Shekhawat, S.R.M. Prasanna, Deep learning techniques for speech emotion recognition: A review, in: 2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA, 2019, pp. 1–6, http://dx.doi.org/10.1109/RADIOELEK.2019.8733432.

[45] A. Singh, P. Rajan, A. Bhavsar, Deep multi-view features from raw audio for acoustic scene classification, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop, DCASE2019, New York University, 2019, pp. 229–233, http://dx.doi.org/10.33682/05gk-pd08, URL http://hdl.handle.net/2451/60765.

[46] J. Huang, H. Lu, P. Lopez-Meyer, H. Cordourier-Maruri, J. Ontiveros, Acoustic scene classification using deep learning-based ensemble averaging, 2019, pp. 94–98, http://dx.doi.org/10.33682/8rd2-g787.

[47] M.H. Bahari, H. Van hamme, Speaker age estimation using hidden Markov model weight supervectors, in: 2012 11th International Conference on Information Science, Signal Processing and their Applications, ISSPA, 2012, pp. 517–521, http://dx.doi.org/10.1109/ISSPA.2012.6310606.

[48] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10 (1) (2000) 19–41, http://dx.doi.org/10.1006/dspr.1999.0361, URL https://www.sciencedirect.com/science/article/pii/S1051200499903615.

[49] T. Bocklet, A. Maier, E. Nöth, Age determination of children in preschool and primary school age with GMM-based supervectors and support vector machines/regression, in: P. Sojka, A. Horák, I. Kopeček, K. Pala (Eds.), Text, Speech and Dialogue, in: Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2008, pp. 253–260, http://dx.doi.org/10.1007/978-3-540-87391-4_33.

[50] T. Bocklet, A. Maier, J.G. Bauer, F. Burkhardt, E. Noth, Age and gender recognition for telephone applications based on GMM supervectors and support vector machines, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 1605–1608, http://dx.doi.org/10.1109/ICASSP.2008.4517932.

[51] S. Prasomphan, Improvement of speech emotion recognition with neural network classifier by using speech spectrogram, in: 2015 International Conference on Systems, Signals and Image Processing, IWSSIP, 2015, pp. 73–76, http://dx.doi.org/10.1109/IWSSIP.2015.7314180.

[52] M. Martinc, F. Haider, S. Pollak, S. Luz, Temporal integration of text transcripts and acoustic features for alzheimer's diagnosis based on spontaneous speech, Front. Aging Neurosci. 13 (2021) http://dx.doi.org/10.3389/fnagi.2021.642647, Cited by: 23; All Open Access, Gold Open Access, Green Open Access.

[53] M.A. Keyvanrad, M.M. Homayounpour, Improvement on automatic speaker gender identification using classifier fusion, in: 2010 18th Iranian Conference on Electrical Engineering, 2010, pp. 538–541, http://dx.doi.org/10.1109/IRANIANCEE.2010.5507010.

[54] Y.M. Ronald Allan Cole, OGI multilanguage corpus - linguistic data consortium, 1994, URL https://catalog.ldc.upenn.edu/LDC94S17.

[55] M. Alsulaiman, Z. Ali, G. Muhammad, Gender classification with voice intensity, in: 2011 UKSim 5th European Symposium on Computer Modeling and Simulation, 2011, pp. 205–209, http://dx.doi.org/10.1109/EMS.2011.37.

[56] A. Pahwa, G. Aggarwal, Speech feature extraction for gender recognition, Int. J. Image, Graph. Signal Process. 8 (2016) 17–25, URL https://api.semanticscholar.org/CorpusID:56291270.

[57] S. Chaudhary, D.K. Sharma, Gender identification based on voice signal characteristics, in: 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN, 2018, pp. 869–874, http://dx.doi.org/10.1109/ICACCCN.2018.8748676.

[58] M.A. Uddin, M.S. Hossain, R.K. Pathan, M. Biswas, Gender recognition from human voice using multi-layer architecture, in: 2020 International Conference on INnovations in Intelligent SysTems and Applications, INISTA, 2020, pp. 1–7, http://dx.doi.org/10.1109/INISTA49547.2020.9194654.

[59] M.A. Uddin, R.K. Pathan, M.S. Hossain, M. Biswas, Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN, J. Inf. Telecommun. 6 (1) (2022) 27–42, http://dx.doi.org/10.1080/24751839.2021.1983318.

[60] M. Büyükyılmaz, A. Çıbıkdiken, Voice gender recognition using deep learning, 2016, http://dx.doi.org/10.2991/msota-16.2016.90.

[61] R. Djemili, H. Bourouba, M.C.A. Korba, A speech signal based gender identification system using four classifiers, in: 2012 International Conference on Multimedia Computing and Systems, 2012, pp. 184–187, URL https://api.semanticscholar.org/CorpusID:17978462.

[62] The IViE Corpus Project Team, The ivie corpus, 2008, URL http://www.phon.ox.ac.uk/files/apps/IViE/.

[63] M.H. Bahari, H. Van Hamme, Speaker age estimation and gender detection based on supervised non-negative matrix factorization, in: 2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications, BIOMS, 2011, pp. 1–6, http://dx.doi.org/10.1109/BIOMS.2011.6052385.

[64] D.A. van Leeuwen, N-best 2008: A benchmark evaluation for large vocabulary speech recognition in dutch, in: P. Spyns, J. Odijk (Eds.), Essential Speech and Language Technology for Dutch: Results By the STEVIN Programme, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 271–288, http://dx.doi.org/10.1007/978-3-642-30910-6_15.

[65] F. Ertam, An effective gender recognition approach using voice data via deeper LSTM networks, Appl. Acoust. 156 (2019) 351–358, http://dx.doi.org/10.1016/j.apacoust.2019.07.033, URL https://www.sciencedirect.com/science/article/pii/S0003682X19304281.

[66] I.E. Livieris, E. Pintelas, P. Pintelas, Gender recognition by voice using an improved self-labeled algorithm, Mach. Learn. Knowl. Extr. 1 (1) (2019) 492–503, http://dx.doi.org/10.3390/make1010030, URL https://www.mdpi.com/2504-4990/1/1/30.

[67] H.A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, M. Rosa-Zurera, Age group classification and gender recognition from speech with temporal convolutional neural networks, Multimedia Tools Appl. (2022) http://dx.doi.org/10.1007/s11042-021-11614-4.

[68] V.S. Kone, A. Anagal, S. Anegundi, P. Jadhav, U. Kulkarni, M.S. M, Voice-based gender and age recognition system, in: 2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT, 2023, pp. 74–80, http://dx.doi.org/10.1109/InCACCT57535.2023.10141801.

[69] Z.-Q. Wang, I. Tashev, Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2017, pp. 5150–5154, http://dx.doi.org/10.1109/ICASSP.2017.7953138.

[70] D. Kwasny, D. Hemmerling, Gender and age estimation methods based on speech using deep neural networks, Sensors 21 (2021) 4785, http://dx.doi.org/10.3390/s21144785.

[71] W.T. Fitch, J. Giedd, Morphology and development of the human vocal tract: A study using magnetic resonance imaging, J. Acoust. Soc. Am. 106 (3) (1999) 1511–1522.

[72] O. Amir, M. Engel, E. Shabtai, N. Amir, Identification of children's gender and age by listeners, J. Voice 26 (3) (2012) 313–321, http://dx.doi.org/10.1016/j.jvoice.2011.06.001, URL https://www.sciencedirect.com/science/article/pii/S0892199711001019.

[73] G. Chen, X. Feng, Y.-L. Shue, A. Alwan, On using voice source measures in automatic gender classification of children's speech, in: Eleventh Annual Conference of the International Speech Communication Association, Citeseer, 2010.

[74] S. Safavi, M. Russell, P. Jančovič, Automatic speaker, age-group and gender identification from children's speech, Comput. Speech & Lang. 50 (2018) 141–156, http://dx.doi.org/10.1016/j.csl.2018.01.001, URL https://www.sciencedirect.com/science/article/pii/S088523081630136X.

[75] M. Guzman, D. Muñoz, M. Vivero, N. Marín, M. Ramírez, M.T. Rivera, C. Vidal, J. Gerhard, C. González, Acoustic markers to differentiate gender in prepubescent children's speaking and singing voice, Int. J. Pediatr. Otorhinolaryngol. 78 (10) (2014) 1592–1598, http://dx.doi.org/10.1016/j.ijporl.2014.06.030, URL https://www.sciencedirect.com/science/article/pii/S0165587614003772.

[76] S.R. Zaman, D. Sadekeen, M.A. Alfaz, R. Shahriyar, One source to detect them all: Gender, age, and emotion detection from voice, in: 2021 IEEE 45th Annual Computers, Software, and Applications Conference, COMPSAC, 2021, pp. 338–343, http://dx.doi.org/10.1109/COMPSAC51774.2021.00055.

[77] C. Müller, Automatic recognition of speakers[2] age and gender on the basis of empirical studies, in: Proc. Interspeech 2006, 2006, http://dx.doi.org/10.21437/Interspeech.2006-195, pp. paper 1031–Wed3CaP.11.

[78] S.B. Kalluri, D. Vijayasenan, S. Ganapathy, Automatic speaker profiling from short duration speech data, Speech Commun. 121 (2020) 16–28, http://dx.doi.org/10.1016/j.specom.2020.03.008, URL https://www.sciencedirect.com/science/article/pii/S0167639319301074.

[79] N. Minematsu, M. Sekiguchi, K. Hirose, Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, IEEE, 2002, pp. I–137.

[80] I. Shafran, M. Riley, M. Mohri, Voice signatures, in: 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721), 2003, pp. 31–36, http://dx.doi.org/10.1109/ASRU.2003.1318399.

[81] A. Gorin, J. Wright, G. Riccardi, A. Abella, T. Alonso, Semantic information processing of spoken language, 2002, URL http://dit.unitn.it/~riccardi/papers/atr2000.pdf.

[82] J. Přibil, A. Přibilová, J. Matoušek, GMM-based speaker gender and age classification after voice conversion, in: 2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines, SPLINE, 2016, pp. 1–5, http://dx.doi.org/10.1109/SPLIM.2016.7528391.

[83] J. Přibil, A. Přibilová, Application of expressive speech in TTS system with cepstral description, in: A. Esposito, N.G. Bourbakis, N. Avouris, I. Hatzilygeroudis (Eds.), Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 200–212.

[84] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-Vectors: Robust DNN embeddings for speaker recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018, pp. 5329–5333, http://dx.doi.org/10.1109/ICASSP.2018.8461375.

[85] L. Wan, Q. Wang, A. Papir, I.L. Moreno, Generalized end-to-end loss for speaker verification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018, pp. 4879–4883, http://dx.doi.org/10.1109/ICASSP.2018.8462665.

[86] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, N. Dehak, Age estimation in short speech utterances based on LSTM recurrent neural networks, IEEE Access 6 (2018) 22524–22530, http://dx.doi.org/10.1109/ACCESS.2018.2816163.

[87] C. Greenberg, A. Martin, D. Graff, L. Brandschain, K. Walker, 2010 NIST speaker recognition evaluation test set - linguistic data consortium, 2010, URL https://catalog.ldc.upenn.edu/LDC2017S06.

[88] S.B. Kalluri, D. Vijayasenan, S. Ganapathy, A deep neural network based end to end model for joint height and age estimation from short duration speech, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019, pp. 6580–6584, http://dx.doi.org/10.1109/ICASSP.2019.8683397.

[89] U.H. Jaid, A.K. Abdulhassan, End-to-end speaker profiling using 1D CNN architectures and filter bank initialization, Int. J. Online Biomed. Eng. (IJOE) (2023) URL https://api.semanticscholar.org/CorpusID:260394711.

[90] S. Rajaa, P.V. Tung, C.E. Siong, Learning speaker representation with semi-supervised learning approach for speaker profiling, 2021, ArXiv abs/2110.13653. URL https://api.semanticscholar.org/CorpusID:239885502.

[91] A. Tursunov, Mustaqeem, J.Y. Choeh, S. Kwon, Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms, Sensors 21 (17) (2021) http://dx.doi.org/10.3390/s21175892, URL https://www.mdpi.com/1424-8220/21/17/5892.

[92] R.B. Lanjewar, S. Mathurkar, N. Patel, Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and K- nearest neighbor (k-NN) techniques, Procedia Comput. Sci. 49 (2015) 50–57, http://dx.doi.org/10.1016/j.procs.2015.04.226, URL https://www.sciencedirect.com/science/article/pii/S1877050915007358. Proceedings of 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15).

[93] A.B. Ingale, D. Chaudhari, Speech emotion recognition, Int. J. Soft Comput. Eng. (IJSCE) 2 (1) (2012) 235–238.

[94] A. Pratama, S.W. Sihwi, Speech emotion recognition model using support vector machine through MFCC audio feature, in: 2022 14th International Conference on Information Technology and Electrical Engineering, ICITEE, 2022, pp. 303–307, URL https://api.semanticscholar.org/CorpusID:253803823.

[95] A. Jacob, P. Mythili, Prosodic feature based speech emotion recognition at segmental and supra segmental levels, in: 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems, SPICES, 2015, pp. 1–5, URL https://api.semanticscholar.org/CorpusID:10044503.

[96] E. Bozkurt, E. Erzin, C.E. Erdem, A.T. Erdem, Use of line spectral frequencies for emotion recognition from speech, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 3708–3711, http://dx.doi.org/10.1109/ICPR.2010.903.

[97] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of german emotional speech, in: Proc. Interspeech 2005, 2005, pp. 1517–1520, http://dx.doi.org/10.21437/Interspeech.2005-446.

[98] D. Technischen, S. Steidl, Automatic classification of emotion-related user states in spontaneous children's speech, 2009, URL https://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2009/Steidl09-ACO.pdf.

[99] C. Parlak, B. Diri, Emotion recognition from the human voice, in: 2013 21st Signal Processing and Communications Applications Conference, SIU 2013, 2013, pp. 1–4, http://dx.doi.org/10.1109/SIU.2013.6531196.

[100] O.E. Korkmaz, A. Atasoy, Emotion recognition from speech signal using mel-frequency cepstral coefficients, in: 2015 9th International Conference on Electrical and Electronics Engineering, ELECO, 2015, pp. 1254–1257, http://dx.doi.org/10.1109/ELECO.2015.7394435.

[101] S. Motamed, S. Setayeshi, A. Rabiee, Speech emotion recognition based on a modified brain emotional learning model, Biol. Inspired Cogn. Archit. 19 (2017) 32–38, http://dx.doi.org/10.1016/j.bica.2016.12.002, URL https://www.sciencedirect.com/science/article/pii/S2212683X16301219.

[102] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, G.-Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280, http://dx.doi.org/10.1016/j.neucom.2017.07.050, URL https://www.sciencedirect.com/science/article/pii/S0925231217313565.

[103] J. Zhang, H. Jia, Design of speech corpus for mandarin text to speech, in: The Blizzard Challenge 2008 Workshop, 2008.

[104] G.H. Mohmad Dar, R. Delhibabu, Speech databases, speech features, and classifiers in speech emotion recognition: A review, IEEE Access 12 (2024) 151122–151152, http://dx.doi.org/10.1109/ACCESS.2024.3476960.

[105] M. Chourasia, S. Haral, S. Bhatkar, S. Kulkarni, Emotion recognition from speech signal using deep learning, in: J. Hemanth, R. Bestak, J.I.-Z. Chen (Eds.), Intelligent Data Communication Technologies and Internet of Things, Springer Singapore, Singapore, 2021, pp. 471–481.

[106] D. Issa, M. Fatih Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomed. Signal Process. Control. 59 (2020) 101894, http://dx.doi.org/10.1016/j.bspc.2020.101894, URL https://www.sciencedirect.com/science/article/pii/S1746809420300501.

[107] Mustaqeem, M. Sajjad, S. Kwon, Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM, IEEE Access 8 (2020) 79861–79875, http://dx.doi.org/10.1109/ACCESS.2020.2990405.

[108] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, http://dx.doi.org/10.1109/5.726791.

[109] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, 2015, arXiv:1411.4038. URL https://arxiv.org/abs/1411.4038.

[110] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, B. Schuller, Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition, IEEE Access 7 (2019) 97515–97525, http://dx.doi.org/10.1109/ACCESS.2019.2928625.

[111] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, K.P. Truong, The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202, http://dx.doi.org/10.1109/TAFFC.2015.2457417.

[112] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM International Conference on Multimedia, MM '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 835–838, http://dx.doi.org/10.1145/2502081.2502224.

[113] T. Anvarjon, Mustaqeem, S. Kwon, Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features, Sensors 20 (18) (2020) http://dx.doi.org/10.3390/s20185212, URL https://www.mdpi.com/1424-8220/20/18/5212.

[114] Mustaqeem, S. Kwon, MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach, Expert Syst. Appl. 167 (2021) 114177, http://dx.doi.org/10.1016/j.eswa.2020.114177, URL https://www.sciencedirect.com/science/article/pii/S0957417420309131.

[115] S. Mishra, N. Bhatnagar, P. Periasamy, S. Nur, Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model, Multimedia Tools Appl. (2023) http://dx.doi.org/10.1007/s11042-023-16849-x.

[116] M.K.P.-F. Kate Dupuis, Toronto emotional speech set (TESS) | tspace repository, 2010, URL https://tspace.library.utoronto.ca/handle/1807/24487.

[117] P. Jackson, S. Haq, Surrey audio-visual expressed emotion (SAVEE) database, 2015, URL http://kahlan.eps.surrey.ac.uk/savee/Download.html.

[118] C. Huang, W. Gong, W. Fu, D. Feng, et al., A research of speech emotion recognition based on deep belief network and SVM, Math. Probl. Eng. 2014 (2014).

[119] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 3687–3691, http://dx.doi.org/10.1109/ICASSP.2013.6638346.

[120] C. Busso, M. Bulut, C.-C. Lee, E.A. Kazemzadeh, E.M. Provost, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359, URL https://api.semanticscholar.org/CorpusID:11820063.

[121] R. Xia, Y. Liu, DBN-ivector framework for acoustic emotion recognition, in: Interspeech, 2016, URL https://api.semanticscholar.org/CorpusID:20805548.

[122] D.H. Ballard, Modular learning in neural networks, in: Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1, AAAI '87, AAAI Press, 1987, pp. 279–284.

[123] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, C. Espy-Wilson, Adversarial auto-encoders for speech based emotion recognition, 2018, arXiv:1806.02146.

[124] M.A. Nicolaou, H. Gunes, M. Pantic, Output-associative RVM regression for dimensional and continuous emotion prediction, Image Vis. Comput. 30 (3) (2012) 186–196, http://dx.doi.org/10.1016/j.imavis.2011.12.005.

[125] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, S. Kollias, Emotion analysis in man-machine interaction systems, in: Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers 1, Springer, 2005, pp. 318–328.

[126] P. Barros, D. Jirak, C. Weber, S. Wermter, Multimodal emotional state recognition using sequence-dependent deep hierarchical features, Neural Netw. 72 (2015) 140–151.

[127] V. Meza-Kubo, A.L. Morán, I. Carrillo, G. Galindo, E. García-Canseco, Assessing the user experience of older adults using a neural network trained to recognize emotions from brain signals, J. Biomed. Inform. 62 (2016) 202–209.

[128] M. Imani, G.A. Montazer, A survey of emotion recognition methods with emphasis on E-learning environments, J. Netw. Comput. Appl. 147 (2019) 102423, http://dx.doi.org/10.1016/j.jnca.2019.102423, URL https://www.sciencedirect.com/science/article/pii/S1084804519302759.

[129] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, M. Pantic, Avec 2011–the first international audio/visual emotion challenge, in: Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II, Springer, 2011, pp. 415–424.

[130] B. Schuller, M. Valster, F. Eyben, R. Cowie, M. Pantic, Avec 2012: the continuous audio/visual emotion challenge, in: Proceedings of the 14th ACM International Conference on Multimodal Interaction, 2012, pp. 449–456.

[131] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, Avec 2013: the continuous audio/visual emotion and depression recognition challenge, in: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 2013, pp. 3–10.

[132] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, M. Pantic, SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild, IEEE Trans. Pattern Anal. Mach. Intell. 43 (3) (2021) 1022–1040, http://dx.doi.org/10.1109/TPAMI.2019.2944808, URL http://arxiv.org/abs/1901.02839. arXiv:1901.02839 [cs].

[133] N.W. Hashim, M. Wilkes, R. Salomon, J. Meggs, D.J. France, Evaluation of voice acoustics as predictors of clinical depression scores, J. Voice 31 (2) (2017) 256.e1–256.e6, http://dx.doi.org/10.1016/j.jvoice.2016.06.006, URL https://www.sciencedirect.com/science/article/pii/S0892199716301059.

[134] M. Hamilton, A rating scale for depression, J. Neurol. Neurosurg. & Psychiatry 23 (1) (1960) 56–62, http://dx.doi.org/10.1136/jnnp.23.1.56, URL https://jnnp.bmj.com/content/23/1/56. Publisher: BMJ Publishing Group Ltd _eprint: https://jnnp.bmj.com/content/23/1/56.full.pdf.

[135] A.T. Beck, C.H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Arch. Gen. Psychiatry 4 (6) (1961) 561–571, http://dx.doi.org/10.1001/archpsyc.1961.01710120031004.

[136] H. Ellgring, K.R. Scherer, Vocal indicators of mood change in depression, J. Nonverbal Behav. 20 (1996) 83–110.

[137] Å. Nilsonne, J. Sundberg, S. Ternström, A. Askenfelt, Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression, J. Acoust. Soc. Am. 83 (2) (1988) 716–728.

[138] E. Moore II, M.A. Clements, J.W. Peifer, L. Weisser, Critical analysis of the impact of glottal features in the classification of clinical depression in speech, IEEE Trans. Biomed. Eng. 55 (1) (2008) 96–107, http://dx.doi.org/10.1109/TBME.2007.900562.

[139] L. Zhang, R. Duvvuri, K.K.L. Chandra, T. Nguyen, R.H. Ghomi, Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative, Depress. Anxiety 37 (2020) 657–669, http://dx.doi.org/10.1002/da.23020.

[140] K.E.B. Ooi, M. Lech, N.B. Allen, Multichannel weighted speech classification system for prediction of major depression in adolescents, IEEE Trans. Biomed. Eng. 60 (2) (2013) 497–506, http://dx.doi.org/10.1109/TBME.2012.2228646.

[141] E. Moore II, M.A. Clements, J.W. Peifer, L. Weisser, Critical analysis of the impact of glottal features in the classification of clinical depression in speech, IEEE Trans. Biomed. Eng. 55 (1) (2007) 96–107.

[142] L.-S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen, Detection of clinical depression in adolescents' speech during family interactions, IEEE Trans. Biomed. Eng. 58 (3) (2010) 574–586.

[143] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk, IEEE Trans. Biomed. Eng. 47 (7) (2000) 829–837.

[144] H. Cai, Z. Yuan, Y. Gao, S. Sun, N. Li, F. Tian, H. Xiao, J. Li, Z. Yang, X. Li, Q. Zhao, Z. Liu, Z. Yao, M. Yang, H. Peng, J. Zhu, X. Zhang, G. Gao, F. Zheng, R. Li, Z. Guo, R. Ma, J. Yang, L. Zhang, X. Hu, Y. Li, B. Hu, A multi-modal open dataset for mental-disorder analysis, Sci. Data 9 (1) (2022) http://dx.doi.org/10.1038/s41597-022-01211-x, URL https://doi.org/10.1038/s41597-022-01211-x.

[145] N. Ghadiri, R. Samani, F. Shahrokh, Integration of text and graph-based features for detecting mental health disorders from voice, 2022, arXiv:2205.07006.

[146] J. Gratch, R. Artstein, G.M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D.R. Traum, A.A. Rizzo, L.-P. Morency, The distress analysis interview corpus of human and computer interviews, in: International Conference on Language Resources and Evaluation, 2014, URL https://api.semanticscholar.org/CorpusID:14488823.

[147] G. Gosztolya, A. Bagi, S. Szalóki, I. Szendi, I. Hoffmann, Identifying Schizophrenia Based on Temporal Parameters in Spontaneous Speech, in: Proc. Interspeech 2018, 2018, pp. 3408–3412, http://dx.doi.org/10.21437/Interspeech.2018-1079.

[148] M. Asgari, L. Chen, E. Fombonne, Quantifying voice characteristics for detecting autism, Front. Psychol. 12 (2021) http://dx.doi.org/10.3389/fpsyg.2021.665096, URL https://www.frontiersin.org/articles/10.3389/fpsyg.2021.665096.

[149] E. Lyakso, O. Frolova, A. Grigorev, A comparison of acoustic features of speech of typically developing children and children with autism spectrum disorders, in: A. Ronzhin, R. Potapova, G. Németh (Eds.), Speech and Computer, Springer International Publishing, Cham, 2016, pp. 43–50.

[150] F. Briend, C. David, S. Silleresi, J. Malvy, S. Ferré, M. Latinus, Voice acoustics allow classifying autism spectrum disorder with high accuracy, Transl. Psychiatry 13 (2023) http://dx.doi.org/10.1038/s41398-023-02554-8.

[151] L.M. De Lau, M.M. Breteler, Epidemiology of parkinson's disease, Lancet Neurol. 5 (6) (2006) 525–535.

[152] T. Pringsheim, N. Jette, A. Frolkis, T.D. Steeves, The prevalence of parkinson's disease: A systematic review and meta-analysis, Mov. Disorders 29 (13) (2014) 1583–1590.

[153] J.R. Orozco-Arroyave, F. Hönig, J.D. Arias-Londoño, J.F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, E. Nöth, Automatic detection of parkinson's disease in running speech spoken in three different languages, J. Acoust. Soc. Am. 139 (1) (2016) 481–500, http://dx.doi.org/10.1121/1.4939739, arXiv:https://pubs.aip.org/asa/jasa/article-pdf/139/1/481/13867953/481_1_online.pdf.

[154] J.R. Orozco-Arroyave, J.D. Arias-Londoño, J.F. Vargas-Bonilla, M.C. Gonzalez-Rátiva, E. Nöth, New spanish speech corpus database for the analysis of people suffering from parkinson's disease, in: LREC, 2014, pp. 342–347.

[155] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, et al., AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition, in: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop, 2018, pp. 3–13.

[156] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, Avec 2016: Depression, mood, and emotion recognition workshop and challenge, in: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 2016, pp. 3–10.

[157] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, et al., AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 3–12.

[158] J. Mekyska, E. Janousova, P. Gomez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, J.B. Alonso-Hernandez, M. Faundez-Zanuy, K.L. de Ipiña, Robust and complex approach of pathological speech signal analysis, Neurocomputing 167 (2015) 94–111, http://dx.doi.org/10.1016/j.neucom.2015.02.085, URL https://www.sciencedirect.com/science/article/pii/S0925231215007304.

[159] P. Kukharchik, D. Martynov, I. Kheidorov, O. Kotov, Vocal fold pathology detection using modified wavelet-like features and support vector machines, in: 2007 15th European Signal Processing Conference, IEEE, 2007, pp. 2214–2218.

[160] P. Harar, J.B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, Z. Smekal, Voice pathology detection using deep learning: a preliminary study, in: 2017 International Conference and Workshop on Bioinspired Intelligence, IWOBI, 2017, pp. 1–4, http://dx.doi.org/10.1109/IWOBI.2017.7985525.

[161] M.A. Mohammed, K.H. Abdulkareem, S.A. Mostafa, M. Khanapi Abd Ghani, M.S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, F.T. AL-Dhief, Voice pathology detection and classification using convolutional neural network model, Appl. Sci. 10 (11) (2020) http://dx.doi.org/10.3390/app10113723, URL https://www.mdpi.com/2076-3417/10/11/3723.

[162] N. Cummins, M. Schmitt, S. Amiriparian, J. Krajewski, B. Schuller, "You sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2017, pp. 3806–3809, http://dx.doi.org/10.1109/EMBC.2017.8037686.

[163] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al., The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism, in: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, 2013.

[164] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J.R. Orozco-Arroyave, E. Nöth, Y. Zhang, F. Weninger, The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition, in: Interspeech, 2015, URL https://api.semanticscholar.org/CorpusID:1514071.

[165] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, Y. Zhang, The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking, in: Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, 2014.

[166] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A.S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, S. Zafeiriou, The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring, Interspeech 2017 (2017) http://dx.doi.org/10.21437/interspeech.2017-43, URL https://www.isca-speech.org/archive_v0/Interspeech_2017/pdfs/0043.PDF.

[167] I. El Emary, M. Fezari, F. Amara, Towards developing a voice pathologies detection system, J. Commun. Technol. Electron. 59 (2014) 1280–1288.

[168] H. Cordeiro, J. Fonseca, I. Guimarães, C. Meneses, Voice pathologies identification speech signals, features and classifiers evaluation, in: 2015 Signal Processing: Algorithms, Architectures, Arrangements, and Applications, SPA, IEEE, 2015, pp. 81–86.

[169] G. Muhammad, M.F. Alhamid, M.S. Hossain, A.S. Almogren, A.V. Vasilakos, Enhanced living by assessing voice pathology using a co-occurrence matrix, Sensors 17 (2) (2017) 267.

[170] I. Hammami, L. Salhi, S. Labidi, Pathological voices detection using support vector machine, in: 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP, IEEE, 2016, pp. 662–666.

[171] C.R. Francis, V.V. Nair, S. Radhika, A scale invariant technique for detection of voice disorders using modified mellin transform, in: 2016 International Conference on Emerging Technological Trends, ICETT, IEEE, 2016, pp. 1–6.

[172] H.-C. Hu, S.-Y. Chang, C.-H. Wang, K.-J. Li, H.-Y. Cho, Y.-T. Chen, C.-J. Lu, T.-P. Tsai, O.K.-S. Lee, Deep learning application for vocal fold disease prediction through voice recognition: Preliminary development study, J. Med. Internet Res. 23 (6) (2021) e25247, http://dx.doi.org/10.2196/25247, URL https://www.jmir.org/2021/6/e25247. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[173] F.T. Al-Dhief, N.M.A. Latiff, N.N.N.A. Malik, N.S. Salim, M.M. Baki, M.A.A. Albadr, M.A. Mohammed, A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms, IEEE Access 8 (2020) 64514–64533, http://dx.doi.org/10.1109/ACCESS.2020.2984925.

[174] J. Miles, S. Hempel, The eysenck personality scales: The eysenck personality questionnaire-revised (EPQ-R) and the eysenck personality profiler (EPP), 2004, URL https://api.semanticscholar.org/CorpusID:227203151.

[175] H. Cattell, A. Mead, The sixteen personality factor questionnaire (16pf), in: The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing, SAGE Publications Ltd, 2008, pp. 135–159, http://dx.doi.org/10.4135/9781849200479.

[176] H.J. Eysenck (Ed.), A Model for Personality, Springer Berlin Heidelberg, 1981, http://dx.doi.org/10.1007/978-3-642-67783-0.

[177] I.B. Myers, L.K. Kirby, K.D. Myers, Introduction to type : A guide to understanding your results on the myers-briggs type indicator, 1980, URL https://api.semanticscholar.org/CorpusID:145675728.

[178] J.M. Digman, Personality structure: Emergence of the five-factor model, Annu. Rev. Psychol. 41 (1) (1990) 417–440.

[179] A.V. Ivanov, G. Riccardi, A.J. Sporka, J. Franc, Recognition of personality traits from human spoken conversations, in: Interspeech, 2011, URL https://api.semanticscholar.org/CorpusID:10380132.

[180] S.I. Levitan, Y. Levitan, G. An, M. Levine, R. Levitan, A. Rosenberg, J. Hirschberg, Identifying individual differences in gender, ethnicity, and personality from dialogue for deception detection, in: Proceedings of the Second Workshop on Computational Approaches to Deception Detection, 2016, URL https://api.semanticscholar.org/CorpusID:14936759.

[181] F. Valente, S. Kim, P. Motlícek, Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus, in: Interspeech, 2012, URL https://api.semanticscholar.org/CorpusID:15958894.

[182] T. Polzehl, S. Möller, F. Metze, Automatically assessing personality from speech, in: 2010 IEEE Fourth International Conference on Semantic Computing, 2010, pp. 134–140, http://dx.doi.org/10.1109/ICSC.2010.41.

[183] F. Weninger, J. Krajewski, A. Batliner, B. Schuller, The voice of leadership: Models and performances of automatic analysis in online speeches, IEEE Trans. Affect. Comput. 3 (4) (2012) 496–508, http://dx.doi.org/10.1109/T-AFFC.2012.15.

[184] N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning-based document modeling for personality detection from text, IEEE Intell. Syst. 32 (2) (2017) 74–79, http://dx.doi.org/10.1109/MIS.2017.23.

[185] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artif. Intell. Rev. 53 (4) (2020) 2313–2339, http://dx.doi.org/10.1007/s10462-019-09770-z.

[186] K.Y. Stanford, S. Mall, N.G. Stanford, Prediction of personality first impressions with deep bimodal LSTM, 2017, URL https://api.semanticscholar.org/CorpusID:2881938.

[187] J. Gorbova, I. Lüsi, A. Litvin, G. Anbarjafari, Automated screening of job candidate based on multimodal video processing, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2017, pp. 1679–1685, http://dx.doi.org/10.1109/CVPRW.2017.214.

[188] R.D.P. Principi, C. Palmero, J.C.S.J. Junior, S. Escalera, On the effect of observed subject biases in apparent personality analysis from audio-visual signals, IEEE Trans. Affect. Comput. 12 (3) (2021) 607–621, http://dx.doi.org/10.1109/TAFFC.2019.2956030.

[189] H.J. Escalante, H. Kaya, A.A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J.C.S.J. Junior, M. Madadi, S. Ayache, E. Viegas, F. Gürpı nar, A.S. Wicaksana, C.C.S. Liem, M.A.J. van Gerven, R. van Lier, Modeling, recognizing, and explaining apparent personality from videos, IEEE Trans. Affect. Comput. 13 (2) (2022) 894–911, http://dx.doi.org/10.1109/TAFFC.2020.2973984.

[190] X. Zhao, Y. Liao, Z. Tang, Y. Xu, X. Tao, D. Wang, G. Wang, H. Lu, Integrating audio and visual modalities for multimodal personality trait recognition via hybrid deep learning, Front. Neurosci. 16 (2023) http://dx.doi.org/10.3389/fnins.2022.1107284, URL https://www.frontiersin.org/articles/10.3389/fnins.2022.1107284.

[191] T. Baltrušaitis, P. Robinson, L.-P. Morency, OpenFace: An open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV, 2016, pp. 1–10, http://dx.doi.org/10.1109/WACV.2016.7477553.

[192] M. Novotny, R. Cmejla, T. Tykalova, Automated prediction of children's age from voice acoustics, Biomed. Signal Process. Control. 81 (2023) 104490, http://dx.doi.org/10.1016/j.bspc.2022.104490, URL https://www.sciencedirect.com/science/article/pii/S1746809422009442.

[193] S. Goyal, V.V. Patage, S. Tiwari, Gender and age group predictions from speech features using multi-layer perceptron model, in: 2020 IEEE 17th India Council International Conference, INDICON, 2020, pp. 1–6, http://dx.doi.org/10.1109/INDICON49873.2020.9342434.

[194] M.H. Bahari, M. McLaren, H. Van hamme, D.A. van Leeuwen, Speaker age estimation using i-vectors, Eng. Appl. Artif. Intell. 34 (2014) 99–108, http://dx.doi.org/10.1016/j.engappai.2014.05.003, URL https://www.sciencedirect.com/science/article/pii/S0952197614001018.

[195] A. Fedorova, O. Glembek, T. Kinnunen, P. Matějka, Exploring ANN backends for i-vector based speaker age estimation, in: Interspeech 2015, ISCA, 2015, pp. 3036–3040, http://dx.doi.org/10.21437/Interspeech.2015-103, URL https://www.isca-archive.org/interspeech_2015/fedorova15_interspeech.html.

[196] V. Nanavare, S. Jagtap, Recognition of human emotions from speech processing, Procedia Comput. Sci. 49 (2015) 24–32, http://dx.doi.org/10.1016/j.procs.2015.04.223, URL https://www.sciencedirect.com/science/article/pii/S1877050915007322. Proceedings of 4th International Conference on Advances in Computing, Communication and Control (ICAC3'15).

[197] C.-C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, Speech Commun. 53 (9) (2011) 1162–1171, http://dx.doi.org/10.1016/j.specom.2011.06.004, URL https://www.sciencedirect.com/science/article/pii/S0167639311000884. Sensing Emotion and Affect - Facing Realism in Speech Processing.

[198] E. Moore, M. Clements, J. Peifer, L. Weisser, Analysis of prosodic variation in speech for clinical depression, in: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439), 3, 2003, pp. 2925–2928 Vol.3, http://dx.doi.org/10.1109/IEMBS.2003.1280531.

[199] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, E.F. Morales, Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients, Pervasive Mob. Comput. 31 (2016) 50–66, http://dx.doi.org/10.1016/j.pmcj.2016.01.008, URL https://www.sciencedirect.com/science/article/pii/S1574119216000109.

[200] P. Kukharchik, D. Martynov, I. Kheidorov, O. Kotov, Vocal fold pathology detection using modified wavelet-like features and support vector machines, in: 2007 15th European Signal Processing Conference, 2007, pp. 2214–2218.

[201] T. Dubuisson, T. Dutoit, B. Gosselin, M. Remacle, On the use of the correlation between acoustic descriptors for the normal/pathological voices discrimination, EURASIP J. Adv. Signal Process. 2009 (1) (2009) 173967, http://dx.doi.org/10.1155/2009/173967, URL https://asp-eurasipjournals.springeropen.com/articles/10.1155/2009/173967.

[202] S. Soni, S. Dey, M.S. Manikandan, Automatic audio event recognition schemes for context-aware audio computing devices, in: 2019 Seventh International Conference on Digital Information Processing and Communications, ICDIPC, 2019, pp. 23–28, http://dx.doi.org/10.1109/ICDIPC.2019.8723713.

[203] J. Singh, R. Joshi, Background sound classification in speech audio segments, in: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019, pp. 1–6, http://dx.doi.org/10.1109/SPED.2019.8906597.

[204] A.V. Ivanov, G. Riccardi, A.J. Sporka, J. Franc, Recognition of personality traits from human spoken conversations, in: Interspeech 2011, ISCA, 2011, pp. 1549–1552, http://dx.doi.org/10.21437/Interspeech.2011-467, URL https://www.isca-archive.org/interspeech_2011/ivanov11_interspeech.html.

[205] F. Valente, S. Kim, P. Motlicek, Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus, in: Proc. Interspeech 2012, 2012, pp. 1183–1186, http://dx.doi.org/10.21437/Interspeech.2012-125.

[206] A.A. Mallouh, Z. Qawaqneh, B.D. Barkana, New transformed features generated by deep bottleneck extractor and a GMM–UBM classifier for speaker age and gender classification, Neural Comput. Appl. 30 (2018) 2581–2593.

[207] M. Markitantov, Transfer learning in speaker's age and gender recognition, in: Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings 22, Springer, 2020, pp. 326–335.

[208] D. Kwasny, D. Hemmerling, Joint gender and age estimation based on speech signals using x-vectors and transfer learning, 2020, arXiv preprint arXiv:2012.01551.

[209] H.A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, M. Rosa-Zurera, Convolutional-recurrent neural network for age and gender prediction from speech, in: 2019 Signal Processing Symposium (SPSympo), IEEE, 2019, pp. 242–245.

[210] S. Rajaa, P. Van Tung, C.E. Siong, Learning speaker representation with semi-supervised learning approach for speaker profiling, 2021, arXiv preprint arXiv:2110.13653.

[211] A. Beke, Forensic speaker profiling in a hungarian speech corpus, in: 2018 9th IEEE International Conference on Cognitive Infocommunications, CogInfoCom, IEEE, 2018, pp. 000379–000384.

[212] A.H. Poorjam, M.H. Bahari, et al., Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals, in: 2014 4th International Conference on Computer and Knowledge Engineering, ICCKE, IEEE, 2014, pp. 7–12.

[213] S. Galgali, S.S. Priyanka, B. Shashank, A.P. Patil, Speaker profiling by extracting paralinguistic parameters using mel frequency cepstral coefficients, in: 2015 International Conference on Applied and Theoretical Computing and Communication Technology, ICATccT, IEEE, 2015, pp. 486–489.

[214] S.A. Fulop, Speech Spectrum Analysis, Springer Science & Business Media, 2011.

[215] Y. Lee, S. Lim, I.-Y. Kwak, CNN-based acoustic scene classification system, Electronics 10 (4) (2021) http://dx.doi.org/10.3390/electronics10040371, URL https://www.mdpi.com/2079-9292/10/4/371.

[216] S. Mun, S. Park, D. Han, H. Ko, Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane, Tech. rep., DCASE2017 Challenge, 2017.

[217] S. Suh, S. Park, Y. Jeong, T. Lee, Designing Acoustic Scene Classification Models with CNN Variants, Tech. rep., DCASE2020 Challenge, 2020.

[218] D. Barchiesi, D. Giannoulis, D. Stowell, M.D. Plumbley, Acoustic scene classification: Classifying environments from the sounds they produce, IEEE Signal Process. Mag. 32 (3) (2015) 16–34, http://dx.doi.org/10.1109/MSP.2014.2326181.

[219] A. Dang, T.H. Vu, J.-C. Wang, A survey of deep learning for polyphonic sound event detection, in: 2017 International Conference on Orange Technologies, ICOT, 2017, pp. 75–78, http://dx.doi.org/10.1109/ICOT.2017.8336092.

[220] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M.D. Plumbley, Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge, IEEE/ ACM Trans. Audio, Speech, Lang. Process. 26 (2) (2018) 379–393, http://dx.doi.org/10.1109/TASLP.2017.2778423.

[221] R. Stiefelhagen, K. Bernardin, R. Bowers, R.T. Rose, M. Michel, J. Garofolo, The CLEAR 2007 evaluation, in: R. Stiefelhagen, R. Bowers, J. Fiscus (Eds.), Multimodal Technologies for Perception of Humans, in: Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2008, pp. 3–34, http://dx.doi.org/10.1007/978-3-540-68585-2_1.

[222] Detection and Classification of Acoustic Scenes and Events, DCASE2023 challenge - DCASE, 2023, URL https://dcase.community/challenge2023/.

[223] Z. Ren, Q. Kong, J. Han, M.D. Plumbley, B.W. Schuller, Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019, pp. 56–60, http://dx.doi.org/10.1109/ICASSP.2019.8683434.

[224] B.W. Schuller, A. Batliner, S. Amiriparian, C. Bergler, M. Gerczuk, N. Holz, P. Larrouy-Maestri, S.P. Bayerl, K. Riedhammer, A. Mallol-Ragolta, M. Pateraki, H. Coppock, I. Kiskin, M. Sinka, S. Roberts, The ACM multimedia 2022 computational paralinguistics challenge: Vocalisations, stuttering, activity, & mosquitoes, 2022, arXiv:2205.06799.

[225] B. Schuller, S. Steidl, A. Batliner, P.B. Marschik, H. Baumeister, F. Dong, S. Hantke, F.B. Pokorny, E.-M. Rathner, K.D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, S. Zafeiriou, The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats, in: Proc. Interspeech 2018, 2018, pp. 122–126, http://dx.doi.org/10.21437/Interspeech.2018-51.

[226] L. Pham, D. Salovic, A. Jalali, A. Schindler, K. Tran, C. Vu, P.X. Nguyen, Robust, general, and low complexity acoustic scene classification systems and an effective visualization for presenting a sound scene context, 2022, arXiv:2210.08610.

[227] K.J. Piczak, ESC: Dataset for environmental sound classification, in: Proceedings of the 23rd ACM International Conference on Multimedia, MM '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1015–1018, http://dx.doi.org/10.1145/2733373.2806390, URL https://doi-org.ezproxy-f.deakin.edu.au/10.1145/2733373.2806390.

[228] L. Nanni, G. Maguolo, S. Brahnam, M. Paci, An ensemble of convolutional neural networks for audio classification, Appl. Sci. 11 (13) (2021) 5796.

[229] P. Rémy, The DARPA TIMIT acoustic-phonetic continuous speech corpus, 2023, URL https://github.com/philipperemy/timit.

[230] N.M.I. Group, 2008 NIST speaker recognition evaluation test set, 2011, http://dx.doi.org/10.35111/fyxw-v682, LDC2011S08.

[231] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: Proceedings of the 12th Conference on Language Resources and Evaluation, LREC 2020, 2020, pp. 4211–4215.

[232] J.H. Hansen, SUSAS, 1999, http://dx.doi.org/10.35111/X4AT-FF87, Artwork Size: 418648 KB Pages: 418648 KB. URL https://catalog.ldc.upenn.edu/LDC99S78.

[233] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG, 2013, pp. 1–8, http://dx.doi.org/10.1109/FG.2013.6553805.

[234] H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, R. Verma, CREMA-D: Crowd-sourced emotional multimodal actors dataset, IEEE Trans. Affect. Comput. 5 (4) (2014) 377–390, http://dx.doi.org/10.1109/TAFFC.2014.2336244.

[235] Y. Li, J. Tao, L. Chao, W. Bao, Y. Liu, CHEAVD: A Chinese natural emotional audio–visual database, J. Ambient. Intell. Humaniz. Comput. 8 (6) (2017) 913–924, http://dx.doi.org/10.1007/s12652-016-0406-z.

[236] S.R. Livingstone, F.A. Russo, The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, PLOS ONE 13 (5) (2018) e0196391, http://dx.doi.org/10.1371/journal.pone.0196391, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391. Publisher: Public Library of Science.

[237] W. Barry, M. Putzer, Saarbrucken voice database, institute of phonetics, Univ. Saarl. (2012).

[238] G. Muhammad, G. Altuwaijri, M. Alsulaiman, Z. Ali, T.A. Mesallam, M. Farahat, K.H. Malki, A. Al-nasheri, Automatic voice pathology detection and classification using vocal tract area irregularity, Biocybern. Biomed. Eng. 36 (2) (2016) 309–317, http://dx.doi.org/10.1016/j.bbe.2016.01.004, URL https://www.sciencedirect.com/science/article/pii/S0208521616300055.

[239] T.A. Mesallam, M. Farahat, K.H. Malki, M. Alsulaiman, Z. Ali, A. Al-nasheri, G. Muhammad, Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms, J. Heal. Eng. 2017 (2017) 1–13, http://dx.doi.org/10.1155/2017/8783751.

[240] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, P. Wellner, The AMI meeting corpus: A pre-announcement, in: S. Renals, S. Bengio (Eds.), Machine Learning for Multimodal Interaction, in: Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2006, pp. 28–39, http://dx.doi.org/10.1007/11677482_3.

[241] H.J. Escalante, I. Guyon, S. Escalera, J. Jacques, M. Madadi, X. Baró, S. Ayache, E. Viegas, Y. Güçlütürk, U. Güçlü, M.A.J. van Gerven, R. van Lier, Design of an explainable machine learning challenge for video interviews, in: 2017 International Joint Conference on Neural Networks, IJCNN, 2017, pp. 3688–3695, http://dx.doi.org/10.1109/IJCNN.2017.7966320.

[242] E. Fonseca, X. Favory, J. Pons, F. Font, X. Serra, FSD50K: An open dataset of human-labeled sound events, IEEE/ ACM Trans. Audio, Speech, Lang. Process. 30 (2022) 829–852.

[243] N. Turpault, R. Serizel, J. Salamon, A.P. Shah, Sound event detection in domestic environments with weakly labeled data and soundscape synthesis, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop, DCASE2019, New York University, 2019, pp. 253–257, http://dx.doi.org/10.33682/006b-jx26, URL http://hdl.handle.net/2451/60771.

[244] A. Mesaros, T. Heittola, T. Virtanen, A multi-device dataset for urban acoustic scene classification, 2018, arXiv:1807.09840.

[245] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2017, pp. 776–780, http://dx.doi.org/10.1109/ICASSP.2017.7952261.

[246] J.W. Kim, C. Yoon, H.Y. Jung, A military audio dataset for situational awareness and surveillance, Sci. Data 11 (1) (2024) 668, http://dx.doi.org/10.1038/s41597-024-03511-w.

[247] J. Salamon, C. Jacoby, J.P. Bello, A dataset and taxonomy for urban sound research, in: Proceedings of the 22nd ACM International Conference on Multimedia, MM '14, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450330633, 2014, pp. 1041–1044, http://dx.doi.org/10.1145/2647868.2655045.

[248] L. Schatzman, A. Strauss, Social class and modes of communication, Am. J. Sociol. 60 (4) (1955) 329–338, http://dx.doi.org/10.1086/221564, URL https://www-journals-uchicago-edu.ezproxy-b.deakin.edu.au/doi/10.1086/221564. Publisher: The University of Chicago Press.

[249] B. Bernstein, Language and social class, Br. J. Sociol. 11 (1960) 271–276, http://dx.doi.org/10.2307/586750, Place: United Kingdom Publisher: Blackwell Publishing.

[250] S.J. Ko, M.S. Sadler, A.D. Galinsky, The sound of power: Conveying and detecting hierarchical rank through voice, Psychol. Sci. 26 (1) (2015) 3–14, http://dx.doi.org/10.1177/0956797614553009, Publisher: SAGE Publications Inc.

[251] F. Kreuk, Y. Adi, M. Cisse, J. Keshet, Fooling end-to-end speaker verification by adversarial examples, 2018, arXiv:1801.03339.

[252] M. Cisse, Y. Adi, N. Neverova, J. Keshet, Houdini: Fooling deep structured prediction models, 2017, arXiv:1707.05373.

[253] W. Brendel, J. Rauber, M. Bethge, Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, 2018, arXiv:1712.04248.

[254] A. Bank, Voice ID - ANZ mobile banking, 2024, URL https://www.anz.com.au/ways-to-bank/mobile-banking-apps/voice-id/. (Accessed 21 March 2025).

[255] S. Staff, Why audio data is the missing link in security strategies, 2023, URL https://www.sdmmag.com/blogs/14-sdm-blog/post/103642-why-audio-data-is-the-missing-link-in-security-strategies. (Accessed 21 March 2025).

[256] R. Singh, Reconstruction of the human persona in 3D from voice, and its reverse, in: R. Singh (Ed.), Profiling Humans from their Voice, Springer, Singapore, 2019, pp. 325–363, http://dx.doi.org/10.1007/978-981-13-8403-5_9.

[257] AltexSoft, Audio analysis: Technologies, application, and examples, 2023, URL https://www.altexsoft.com/blog/audio-analysis/. (Accessed 21 March 2025).

[258] S. Spiekermann-Hoff, Networks of Control – A Report on Corporate Surveillance, Digital Tracking, Facultas Verlags- und Buchhandels AG, Austria, 2016.

[259] A. Nagrani, J.S. Chung, A. Zisserman, Voxceleb: A large-scale speaker identification dataset, 2017, arXiv preprint arXiv:1706.08612.

[260] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, N. Zeghidour, AudioLM: A language modeling approach to audio generation, IEEE/ ACM Trans. Audio, Speech, Lang. Process. 31 (2023) 2523–2533, http://dx.doi.org/10.1109/TASLP.2023.3288409, URL https://ieeexplore.ieee.org/document/10158503/.

[261] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter, xLSTM: Extended long short-term memory, 2024, arXiv:2405.04517. URL https://arxiv.org/abs/2405.04517.

[262] The elevenLabs Team, Text to Speech & AI Voice Generator, URL https://elevenlabs.io.

[263] J. Chorowski, R.J. Weiss, S. Bengio, A. van den Oord, Unsupervised speech representation learning using WaveNet autoencoders, IEEE/ ACM Trans. Audio, Speech, Lang. Process. 27 (12) (2019) 2041–2053, http://dx.doi.org/10.1109/TASLP.2019.2938863.

[264] X. Wang, S. Takaki, J. Yamagishi, S. King, K. Tokuda, A vector quantized variational autoencoder (VQ-VAE) autoregressive neural F_0 model for statistical parametric speech synthesis, IEEE/ ACM Trans. Audio, Speech, Lang. Process. 28 (2020) 157–170, http://dx.doi.org/10.1109/TASLP.2019.2950099.

[265] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noé, J.-F. Bonastre, M. Todisco, N. Evans, The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment, in: Proc. Interspeech 2020, 2020, pp. 1698–1702, http://dx.doi.org/10.21437/Interspeech.2020-1815.

[266] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, J.-F. Bonastre, Adversarial disentanglement of speaker representation for attribute-driven privacy preservation, 2021, arXiv:2012.04454.

[267] A. Nelus, R. Martin, Privacy-aware feature extraction for gender discrimination versus speaker identification, 2019, pp. 671–674, http://dx.doi.org/10.1109/ICASSP.2019.8682394.

[268] R. Aloufi, H. Haddadi, D. Boyle, Emotionless: Privacy-preserving speech analysis for voice assistants, 2019, URL http://arxiv.org/abs/1908.03632. arXiv:1908.03632 [cs, eess, stat].

[269] D. Stoidis, A. Cavallaro, Generating gender-ambiguous voices for privacy-preserving speech recognition, in: Interspeech 2022, ISCA, 2022, pp. 4237–4241, http://dx.doi.org/10.21437/Interspeech.2022-11322, URL https://www.isca-archive.org/interspeech_2022/stoidis22_interspeech.html.

[270] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2015, pp. 5206–5210, http://dx.doi.org/10.1109/ICASSP.2015.7178964.

[271] D. Stoidis, A. Cavallaro, Protecting gender and identity with disentangled speech representations, 2021, URL http://arxiv.org/abs/2104.11051. arXiv:2104.11051 [cs, eess].

[272] D. Ericsson, A. Östberg, E.L. Zec, J. Martinsson, O. Mogren, Adversarial representation learning for private speech generation, 2020, arXiv:2006.09114. URL https://arxiv.org/abs/2006.09114.

[273] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, H. Lan, F. Zhao, L. Xiong, Y. Xu, J. Li, S. Pranata, S. Shen, J. Xing, H. Liu, S. Yan, J. Feng, Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition, 2018, arXiv:1809.00338. URL https://arxiv.org/abs/1809.00338.

[274] C. Eom, B. Ham, Learning disentangled representation for robust person re-identification, in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, URL https://proceedings.neurips.cc/paper/2019/hash/d3aeec875c479e55d1cdeea161842ec6-Abstract.html.

[275] Y.-H. Tsai, K. Sohn, S. Schulter, M. Chandraker, Domain adaptation for structured output via discriminative patch representations, 2019, arXiv:1901.05427. URL https://arxiv.org/abs/1901.05427.

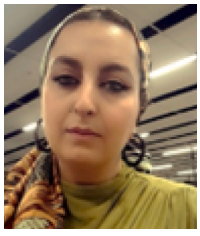
[276] R. Aloufi, H. Haddadi, D. Boyle, Privacy-preserving voice analysis via disentangled representations, in: Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, CCSW '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–14, http://dx.doi.org/10.1145/3411495.3421355, URL https://dl.acm.org/doi/10.1145/3411495.3421355.

[277] R. Aloufi, H. Haddadi, D. Boyle, Paralinguistic privacy protection at the edge, 2022, http://dx.doi.org/10.48550/arXiv.2011.02930, URL http://arxiv.org/abs/2011.02930. arXiv:2011.02930 [cs].

[278] A. van den Oord, O. Vinyals, K. Kavukcuoglu, Neural discrete representation learning, 2018, arXiv:1711.00937. URL https://arxiv.org/abs/1711.00937.
[279] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, Y. Deng, VoiceMask: Anonymize and sanitize voice input on mobile devices, 2017, URL http://arxiv.org/abs/1711.11460. arXiv:1711.11460 [cs].
[280] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, N. Evans, Speaker anonymisation using the McAdams coefficient, in: Proc. Interspeech 2021, 2021, pp. 1099–1103, http://dx.doi.org/10.21437/Interspeech.2021-1070.
[281] S. Zhang, Z. Li, A. Das, VoicePM: A robust privacy measurement on voice anonymity, in: Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks, ACM, Guildford United Kingdom, 2023, pp. 215–226, http://dx.doi.org/10.1145/3558482.3590175, URL https://dl.acm.org/doi/10.1145/3558482.3590175.
[282] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, M. Todisco, The VoicePrivacy 2024 challenge evaluation plan, 2024, URL http://arxiv.org/abs/2404.02677. arXiv:2404.02677 [cs, eess].
[283] O.d. Chouchane, M. Panariello, O. Zari, I. Kerenciler, I. Chihaoui, M. Todisco, M. Önen, Differentially private adversarial auto-encoder to protect gender in voice biometrics, in: Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security, ACM, Chicago IL USA, 2023, pp. 127–132, http://dx.doi.org/10.1145/3577163.3595102, URL https://dl.acm.org/doi/10.1145/3577163.3595102.
[284] M. Panariello, F. Nespoli, M. Todisco, N. Evans, Speaker anonymization using neural audio codec language models, 2024, arXiv:2309.14129. URL https://arxiv.org/abs/2309.14129.
[285] A. Défossez, J. Copet, G. Synnaeve, Y. Adi, High fidelity neural audio compression, 2022, arXiv:2210.13438. URL https://arxiv.org/abs/2210.13438.
[286] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, F. Wei, Neural codec language models are zero-shot text to speech synthesizers, 2023, arXiv:2301.02111. URL https://arxiv.org/abs/2301.02111.

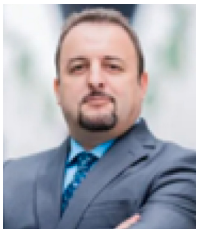
Anil Pudasaini is a Ph.D. student and researcher at Deakin University, Australia. He holds a master's degree in information and communication engineering and a bachelor's degree in computer engineering from Tribhuvan University, Nepal. Prior to joining doctoral studies, he worked as a lecturer, researcher, and project supervisor at various engineering colleges in Nepal.

His research interests lie in the areas of machine learning, deep learning, and data science. He has published five research papers in reputed journals and conference proceedings, covering topics such as sentiment analysis of social media data, concrete compressive strength prediction using machine learning, and digital music generation using LSTMs.

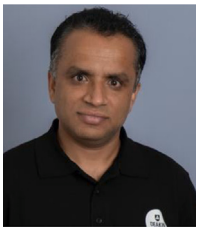


Muna Al-Hawawreh is an Assistant Professor (Lecturer) at the School of Information Technology at Deakin University, Australia. She received her Ph.D. degree in Computer Science from UNSW Canberra in 2022. She also received bachelor's and master's degrees (First Class Honors) in computer science from Mutah University, Jordan. Dr Al-Hawawreh's research is multidisciplinary and focuses on cyber security and privacy-preserving in cyber environments like the industrial Internet of things, industrial control systems, cloud computing, cyber–physical systems, and smart satellites, with a focus on investigating, analyzing, and detecting current and future cyber-attacks (offensive and defensive research). She is also looking into using artificial intelligence applications for cybersecurity automation. She has a strong publication record and has published many peer-reviewed research papers in top-tier journals. Her contribution is recognized both nationally and internationally through achieving various awards.

Mohamed Reda Bouadjenek is a Senior Lecturer (Assistant Professor) of Applied Artificial Intelligence at the School of Information Technology at Deakin University, Australia. He currently co-leads the Machine Learning for Decision Support (MLDS) Group at Deakin University, Geelong Waurn Ponds campus. Before joining Deakin University, he was a Research Fellow at the University of Toronto (2017–2019) and at the University of Melbourne (2015–2017). Prior to that, he was a postdoctoral researcher at INRIA in France (2014–2015). He earned his Ph.D. and M.Sc. in Computer Science from the University of Paris-Saclay, France, in 2013 and 2009, respectively, and a B.Sc. in Computer Science from USTHB, Algeria, in 2008. His research focuses on developing efficient algorithms to extract valuable knowledge from data. He explores a broad range of topics related to machine learning, deep learning, and information retrieval. He publishes his research in top-tier venus like ACM SIGIR, CIKM, IEEE TNNLS, WWW, and Pattern Recognition.



Hakim Hacid is the Chief Researcher of the Artificial Intelligence and Digital Science Research Center at the Technology Innovation Institute (TII), UAE. He obtained his Ph.D. from the University of Lyon (France) in 2007 and worked at the University of New South Wales (Australia), Bell Labs (France) and Zayed University (UAE) at various capacities before joining TII in 2022. He is an honorary professor at Macquarie University, Australia. His research interests include data analytics (structured and unstructured), big data, and web information systems. Over the years, he has been involved in building research projects and large collaborations. He has published more than 80 research articles in top journals and conferences, as well as several industrial patents. Dr. Hacid is also highly active in the international research community, organizing conferences and reviewing research work and research proposals.



Sunil Aryal is an Associate Professor of data science at the School of IT at Deakin University Australia, where he co-lead the Machine Learning for Decision Support (MLDS) Research Group. His research interests are in the areas of Artificial Intelligence (AI), Machine Learning (ML) and Data Mining (DM). He is working on anomaly detection, clustering, kernel/similarity-based learning, ensemble methods, learning from heterogeneous, noisy and limited/small data, reinforcement learning, natural language processing, and computer vision. He is particularly interested in applying AI/ML/DM models to solve real-world problems in Defence, Cybersecurity, Engineering, Agriculture, Healthcare and Education. He has published over 75 papers in top-tier international venues and secured over AUD 4.5M in external research funding from agencies like the US and Australian Defence, State Department of Education and Training Victoria, Technology Innovation Institute UAE, and Table Tennis Australia.