ISSN: 0067-2904

Review of Automatic Speaker Profiling: Features, Methods, and Challenges

Umniah Hameed Jaid*[1], Alia Karim Abdul Hassan[2]

[1]*Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq*
[2] *Department of Computer Science, University of Technology, Baghdad, Iraq*

Received: 4/9/2022 Accepted: 30/1/2023 Published: 30/12/2023

Abstract

Automatic Speaker Profiling (ASP), is concerned with estimating the physical traits of a person from their voice. These traits include gender, age, ethnicity, and physical parameters. Reliable ASP has a wide range of applications such as mobile shopping, customer service, robotics, forensics, security, and surveillance systems. Research in ASP has gained interest in the last decade, however, it was focused on different tasks individually, such as age, height, or gender. In this work, a review of existing studies on different tasks of speaker profiling is performed. These tasks include age estimation and classification, gender detection, height, and weight estimation This study aims to provide insight into the work of ASP, available datasets, feature extraction techniques, and learning models. Further, the performance of current speaker profiling systems is investigated. Finally, the challenges of speaker profiling are presented at the end of this review.

Keywords: Automatic speaker profiling, feature extraction, age estimation, height estimation, estimation, and gender detection.

مراجعة للتنميط التلقائي للمتحدث: الخصائص, الطرق, و التحديات

أمنية حميد جاعد [1]*, علياء كريم عبد الحسن[2]

[1]قسم علوم الحاسوب, كلية العلوم, جامعة بغداد, بغداد, العراق
[2]علوم الحاسوب, الجامعة التكنولوجية, بغداد, العراق

الخلاصة

التنميط التلقائي للمتحدث يهتم بتقدير السمات الجسدية للشخص من صوته. تشمل هذه السمات الجنس والعمر والعرق والمعلمات الفيزيائية تشمل تطبيقات التنميط التلقائي من الصوت مجموعة واسعة من التطبيقات التي تمتد من المجالات التي تتطلب تفاعلًا بين الإنسان والحاسوب، مثل التسوق عبر الهاتف المحمول وخدمة العملاء والروبوتات إلى أنظمة الادلة الجنائية والأمن والمراقبة. وقد اكتسب البحث في هذا المجال اهتمامًا في العقد الماضي، ومع ذلك، يركز البحث على مهام مختلفة بشكل فردي، مثل تقدير العمر أو الطول أو الجنس. في هذا العمل، يتم إجراء مراجعة للدراسات الحالية حول المهام المختلفة لتحديد سمات المتحدثين، وتشمل هذه المهام تقدير العمر وتصنيفه، واكتشاف الجنس، والطول، وتقدير الوزن. على حد علمنا، لا يوجد استبيان حول هذا الموضوع. ومن ثم تهدف هذه الدراسة إلى توفير نظرة مدخل الى طرق استخراج هذه السمات والبيانات

*Email: umniah.h@sc.uobaghdad.edu.iq

المتاحة وتقنيات استخراج الميزات ونماذج التعلم. علاوة على ذلك، يتم التحقيق في أداء أنظمة التنميط الحالية
للمتحدث. أخيرًا، يتم تقديم التحديات الي تواجه الباحثين في هذا المجال.

1. Introduction

Speech is one of the main means of communication between humans, as they express their thoughts, ideas, information, and emotions through speech.

A speech signal can convey far more information beyond the spoken words. The human ear can easily recognize the gender of the speaker, the age group of the speaker (senior, youth, or child), and the emotional and physical state of the speaker from their voice (tired, stressed, relaxed).

Speaker profiling is a process that humans perform seamlessly and unconsciously. A person immediately recognizes the gender and age group of a speaker and can estimate the speaker’s physical shape and emotions from the sound of their voice and adapt responses accordingly.

Automated Speaker Profiling (ASP) refers broadly to the computational extraction/ prediction of speaker traits from the speech signal. The predicted traits may be Physical characteristics (e.g., gender, height, weight, age), Psychological (e.g., emotional state, stress level), Social (e.g., ethnicity, education level, socio-economic status, dialect-region), or Biomarker (voice data used for detecting alterations in health).
Applications of ASP span many fields and have a wide range of real-world applications in surveillance [1], forensics [2, 3], commercial [4, 5], and Human-robot interaction [6, 7] (Figure 1).

In particular, voice offers major advantages over the other biometric processes in terms of intrusiveness, cost, ease of deployment, and user acceptance, since it is the least intrusive among the many biometrics being used[8].

The ability to automatically extract a speaker’s profile would support the quick evaluation of expansive volumes of recordings, for which manual triage may not be attainable[9]. Moreover, ASP can be used to augment data from surveillance systems[1]. For example, a camera in a user identification system may have a partial view of the user, or the image of the user is occluded by another object. Similarly, Speaker profiling can be used in forensics for obtaining a description of an unknown speaker [2, 3]. Commercial uses of ASP include call routing based on the speaker profile, playing appropriate music/messages in call waiting, assigning appropriate customer service to the caller, and mobile shopping [4].

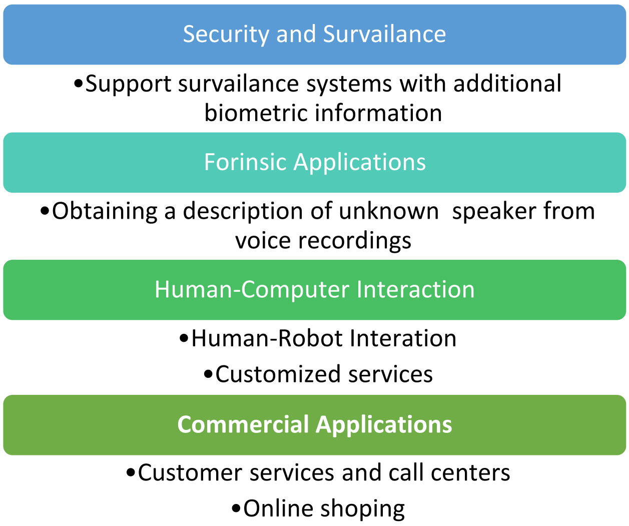
Security and Survailance
•Support survailance systems with additional biometric information
Forinsic Applications
•Obtaining a description of unknown speaker from voice recordings
Human-Computer Interaction
•Human-Robot Interation
•Customized services
Commercial Applications
•Customer services and call centers
•Online shoping

Figure 1: Applications of Automatic Speaker Profiling

Existing work in speaker profiling can be broadly classified into two categories as shown in Figure 2. The first category is single-task profiling, where the ASP system concentrates on estimating only one physical parameter at a time, such as age prediction or classification into an age group, height estimation, and gender detection. The second category is multi-task profiling, in which several traits are estimated using the same feature vector and learning model.

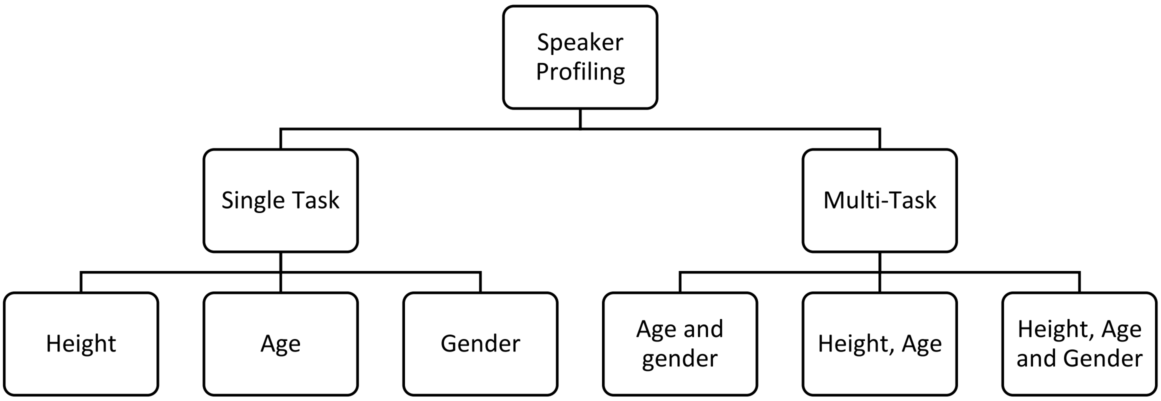

Figure 2: Types of Speaker profiling in reviewed articles

Age estimation and classification are well studied in the literature, however, detecting age is considered one of the most challenging tasks in ASP. The challenge arises from the overlap of different sources of variability, where speech is affected by many factors other than age, such as the shape of the sound production system, gender, health, and emotional state of the speaker. However, some characteristics of the voice can indicate the age of the speaker; these characteristics include the speech rate, as younger speakers have a higher speech rate than older speakers[4]. The fundamental frequency is also an indication of age, it tends to decrease with age [9], especially for female speakers [10]. Other characteristics, such as jitter and shimmer were also found to be affected by age [11].

Gender detection in ASP literature is often coupled with age group classification, and it is mainly performed with the help of the fundamental frequency (f0), as male speakers usually have lower frequencies than female speakers. However, in children, identifying gender is more challenging, as the sex differences in vocal tract structure are not present in children and don't appear until puberty[12].

Height estimation of a speaker from their speech can be largely attributed to the assumption that the length of the vocal tract (VTL) is directly proportional to the speaker's height [13]. Several studies also showed the ability of listeners to accurately estimate the relative size of the speaker from their voice (i.e., height and weight). Furthermore, a study involving magnetic resonance imaging (MRI) of 129 individuals confirmed the correlation between a person's height and weight and their vocal tract length (VTL) [12]. Research on weight estimation is very limited due to the lack of available data on speakers' weight. Weight estimation is often coupled with height estimation [14, 15].

In General, a speaker profiling system in the literature can be divided into three parts as can be seen in Figure 3, Data Acquisition and Pre-processing, Feature Extraction and Representation, and a learning model. These parts will be discussed in the following sections. A list of abbreviations used in this work with their full form is presented in Table 1.

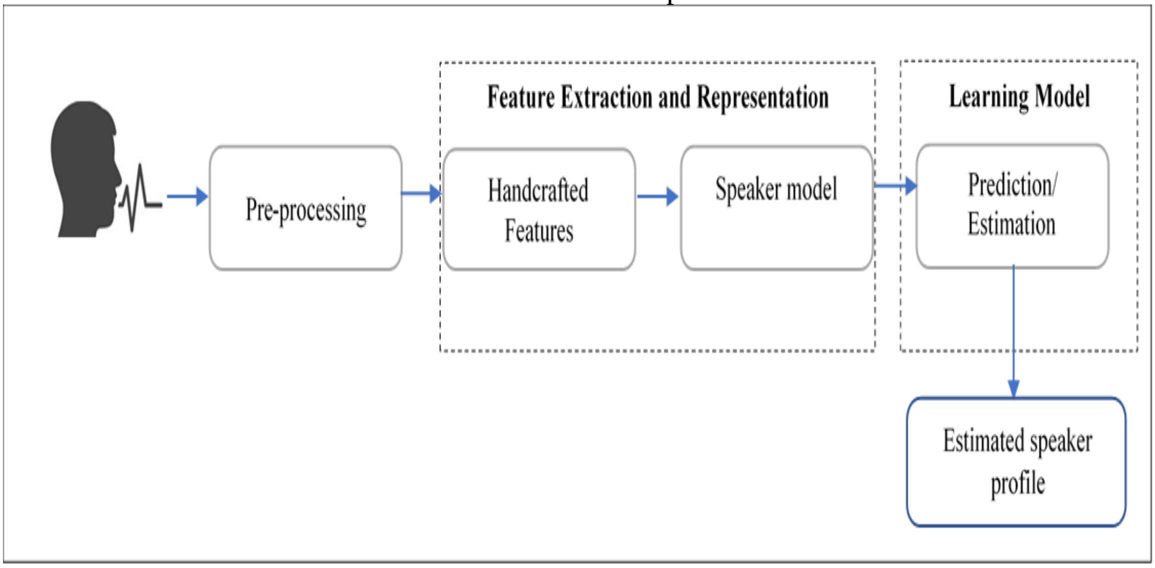

Figure 1: General structure of automatic speaker profiling process model.

In this work, we present a comprehensive review of speaker profiling literature. This survey aims to discuss the findings of different related research on speaker profiling, such as estimation of a speaker's gender, age, height, and weight.

The aspects of the review include the most used databases for profiling, common feature extraction methods, and various machine learning and deep learning models employed for ASP. Lastly, the paper outlines the current challenges and presents recommendations for future work in ASP. The contributions of this paper are outlined as follows:

- Overview of the state of the art and the current research direction in speaker profiling.
- Provide the reader with a list of available speaker profiling databases, common feature extraction methods, and learning models used for different aspects of ASP.
- Discuss strength, challenges, and shortcomings of the current practices in ASP.

The remainder of this study is arranged as follows: Section 2 reviews available datasets for ASP and their properties. Section 3 provides a general overview of the steps involved in signal pre-processing, while Section 4 outlines feature extraction methods that appear in the ASP literature. Section 5 examines the learning models applied in ASP. Section 6 provides a discussion of the findings from reviewed work, and Section 7 describes current challenges in ASP research. Finally, Section 8 concludes this review article.

Table 1: List of Abbreviations mentioned in this work

Abbreviations	Definition
ASP	Automatic Speaker Profiling
MTL	Multi-task Learning
VTL	Vocal Tract Length
MFCC	Mel Frequency Cepstral Coefficients
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral Coefficients
ZCR	Zero Crossing Rate
RMS	Root Mean Square
F0	Fundamental Frequency
HNR	Harmonic Noise Ratio
UBM	Universal Background Model
GMM	Gaussian Mixture Model
LSF	Line Spectral Frequencies
SCC	Spectral Sub band Centroid
DFT	Discrete Fourier Transform
DCT	Discrete Cosine Transform
DNN	Deep Neural Network
CNN	Convolutional Neural Networks
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
UA	Unweighted Accuracy
LSTM	Long Short Term Memory
TDNN	Time-Delay Neural Network
SVM	Support Vector Machine
SVR	Support Vector Regression
LSSVR	Least Square SVR
ANN	Artificial Neural Networks
LR	Linear Regression
DT	Decision Trees

2. Datasets

In this section, several databases employed in ASP literature are discussed. As mentioned above, speaker profiling can be divided into several tasks. These tasks can be assessed individually or in combination. Available speech datasets vary in the level of details they contain about the individuals in the recordings. Figure 4 Presents the frequency of most-used datasets for speaker profiling in the literature. The most widely used dataset is the TIMIT speech corpus [16]. This corpus consists of speech utterances of 630 speakers, 70% of which are male speakers from eight major dialect regions (New York City, New England, Northern, South Midland, North Midland, Southern, Army Brat, Western) of the USA. Each speaker has 10 recordings of different phonetically rich sentences, which makes this dataset also suitable for speech recognition tasks.

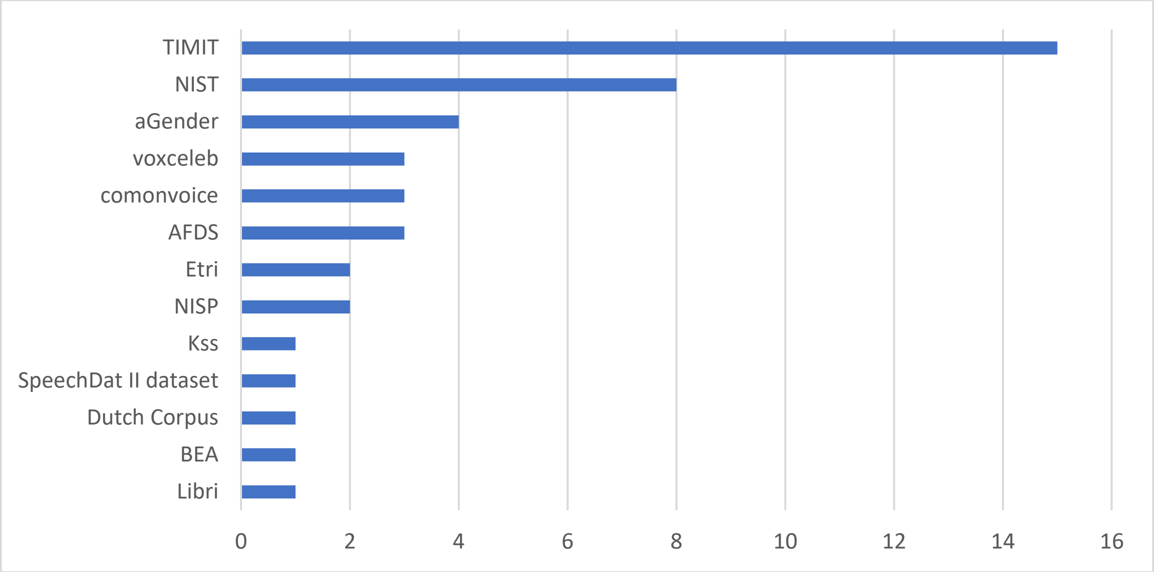

Figure 2: Occurrence of speaker profiling datasets in the studies considered in this review

Table 2: summary of speech datasets used in reviewed articles

Dataset	*Language*	*Access*	*Size / No. of Speakers*	*Gender*	*Age*	*Height*	*Weight*	*Other*
TIMIT	English	Paid	630	✓	✓	✓		✓
NISP	multi-language	Free[1]	345	✓	✓	✓	✓	✓
NIST-SRE 2008	multi-language	Paid[2]	1236	✓	✓	✓	✓	✓
NIST-SRE 2010	multi-language	Paid[2]	445					
Voxceleb	English	Free[3]	153,516	✓				✓
AFDS	multi-language	Private	207	✓		✓	✓	✓
aGender	German	Paid[4]	954	✓	✓			
Common voice	multi-language	Free[5]	14,122 hours	✓	✓			
Librispeech	English	Free[6]	251	✓				
BEA	Hungarian	Private[7]	280	✓	✓	✓	✓	
KSS DB	Korean	Private	2500	✓	✓			
Dutch Corpus	Dutch	Private	1,000 hours	✓	✓			
SpeechDat II dataset	German	paid[8]	4000	✓	✓			
ETRI-VoiceDB2006	NA	Private	58	✓	✓			

[1] https://github.com/iiscleap/NISP-Dataset

[2] https://sre.nist.gov/

[3] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

[4] https://www.isca-speech.org/iscapad/iscapad.php?module=category&id=684

[5] https://voice.mozilla.org/en/datasets

[6] https://www.openslr.org/12

[7] http://www.nytud.hu/dbases/bea/index.html

[8] https://catalogue.elra.info/en-us/repository/browse/ELRA-S0105/

Additionally, the dataset includes information about the speakers such as age, height, race, and level of education.

Another commonly used dataset for profiling is the NIST SRE 2008 and 2010 database [17]. The National Institute of Standards and Technology (NIST) conducts yearly or biannual speaker recognition evaluations (SRE). Each SRE includes a huge corpus of phone (and, more lately, microphone) conversations. These chats usually take 5 minutes and involve a large number of speakers for whom additional meta-information like age, height, weight, language, and smoking habits is collected.

In addition, there is a recently published dataset by kalluri et. al.[18], named NITK-IISc Multilingual Multi-accent Speaker Profiling or NISP dataset collected specifically for the task of speaker profiling. There are 345 speakers in the NISP dataset, including 219 males and 126 females. Five native Indian languages are included in the dataset, as well as Indian-accented English. In each language, every speaker supplied around 4-5 minutes of speech data. The Metadata in NISP includes Physical characteristics such as age, gender, height, shoulder size, and weight among other linguistic information.

The aforementioned datasets are suitable for multi-task ASP, however, several other less used datasets are employed for age and gender classification, such as aGender [19], common voice [20], and VoxCeleb [21]. A summary of datasets that appear in speaker profiling literature is presented in table 2.

There are several factors to be considered when choosing a data set for ASP, these factors are discussed in the following sections.

2.1 Data Imbalance

A key factor in the performance of an ASP system is having enough data to achieve acceptable results. However, the balance between age groups and male / female recordings is not often presented. For instance, the TIMIT datasets include recordings of 630 speakers, only 192 of them are females while the rest (438) are male speakers. Similarly, the VoxCeleb dataset contains 1251 speakers, the data is more balanced as it contains 55% male speakers to 45% female speakers. A list of datasets employed for gender detection in reviewed articles and the distribution of male to female speakers is presented in Figure 5. Aside from gender imbalance, another issue with these datasets is the limited range of ages, and the imbalance of age groups represented amongst the speakers.

2.2 Quality of Recordings

The performance of an ASP system may be affected by the quality of the recordings in terms of noise or duration of the recordings. Recordings durations vary by different datasets, for example, NIST-SRE 2008/2010 includes recordings as long as 10 minutes. While TIMIT dataset recording durations vary from 2-5 seconds. A number of studies evaluate the effect of utterance duration on ASP tasks [11] . Additionally, environment and ambient noise can influence the performance of ASP. Some datasets are collected under controlled environments, while other datasets are recorded in different environments, such as VoxCeleb.

2.3 Limited Data

Most datasets include gender and age information about speakers. Thus, finding data for gender detection and age estimation tasks is not a challenge. However, datasets for estimating

Height and Weight are limited. The TIMIT dataset includes the height of each speaker. While the NISP dataset contains additional data about weight, shoulder size, neck, and waist. The NIST SRE 2008 and 2010 databases that are commonly used for speech recognition also include data about the weight and height of the speakers.
Additionally, the Hungarian BAE dataset includes information about the height and weight of speakers in addition to age and gender.

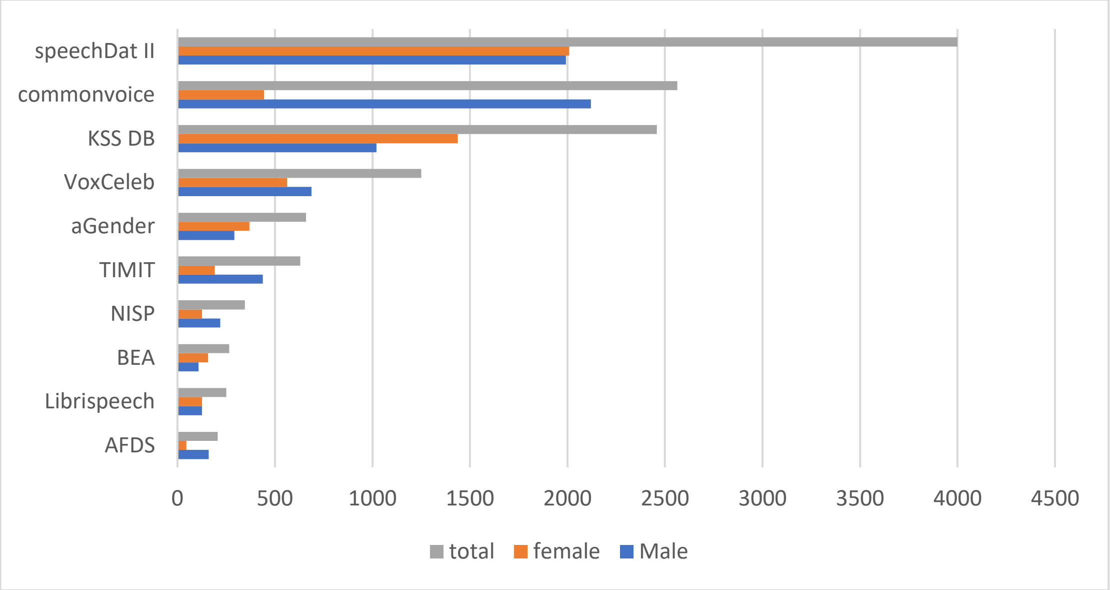

Figure 3: Gender distribution in ASP dataset

3. Pre-processing

Before processing the speech signal and starting the feature extraction process, [22] pre-processing and segmentation of the speech signal is required. Pre-processing steps include pre-emphasis, framing, and windowing. Figure 6 illustrates these steps and their effect on the signal.

Pre-emphasis: one of the first steps in pre-processing where the magnitude of higher frequencies is increased compared to lower frequencies, and it is often implemented as:

$$y_t = x_t - \alpha x_{t-1} \qquad (1)$$

Where x_t is the sample signal at time t, x_{t-1} is the sample of the previous time step, α is the weight factor, and y_t is the emphasised sample at time t.
Pre-emphasis improves the overall signal-to-noise ratio by highlighting the more discriminative higher frequencies [22].

Framing: Signal processing assumes that the signal is stationary; however, the speech signal is constantly and slowly varying over time. Thus, the spectral features need to be extracted over a sufficiently short period or frames where the signal is assumed to remain stationary, and a spectral feature vector is extracted from each frame. In general, short-term spectral features are computed from a frame size of about 20-30ms in duration with an overlap of half of the frame length.

Windowing: The process of segmenting the signal into frames causes discontinuities at the edges of the segments, which in turn leads to spectral leakage. To alleviate this effect, a tapered window function is applied. The Hamming window is one of the most common approaches[23]. A Hamming window has the following form:

$$w[n] = 0.54 - 0.46 \cos [(\frac{2\pi n}{N-1})] \qquad (2)$$

Where N is the length of the frame, and n = 0, 1, …, N – 1.

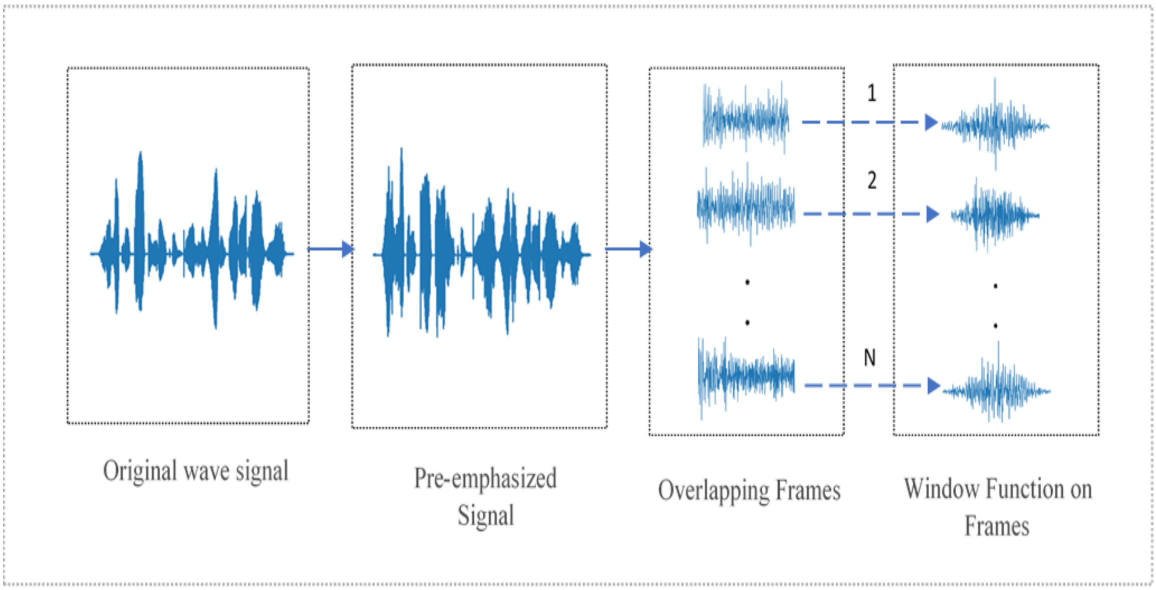

Figure 4: Visualization of the steps of audio signal pre-processing

4. Feature Extraction Approaches in ASP

One of the main concerns of any speech processing system is the representation of variable-size utterances with a fixed-size feature vector. The first phase in the ASP process is concerned with the extraction and representation of discriminative features from speech utterances to be used in model training and prediction. That involves converting time-domain raw speech samples into compact and efficient feature vectors that retain the speaker information. Features employed in ASP can be broadly categorized into handcrafted features and learned features. Handcrafted features are directly extracted from the speech utterance, such as formants and the fundamental frequency. On the other hand, learned features are obtained by training a model on these extracted features to obtain a low-level speaker representation or speaker models, such as supervectors and i-vectors.

The feature extraction methods employed in ASP are discussed in this section, and Table 3 provides a summary of features employed in the reviewed articles.

Table 3 : summary of features employed in the reviewed articles sorted according to the publishing year

Reference	Authors/Year	Task	Features	Dataset
[24]	Dusan, 2005	Height	MFCC+ LPC+F0+Formnats	TIMIT
[5]	Müller & Burkhardt, 2007	Age	MFCC+Pitch	SpeechDat II
[1]	Mporas & Ganchev, 2010	Height	ZCR+RMS+ Frame Energy+F0+HNR+MFCC+Mean+std+ kurtosis+ Skewness+ minimum + maximum	TIMIT
[25]	Bahari & Van Hamme, 2011	Age	GMM supervector(Mel spectrogram)	Dutch Corpus
[26]	Bahari, 2012	Age	HMM supervector(Mel spectrogram)	Dutch Corpus
[27]	Williams & Hansen, 2013	Height	Formants + LPC	TIMIT
[2]	Poorjam & Bahari, 2014	Height, weight, smoking habits	i-vectors(MFCC)	NIST
[28]	Arsikere et al., 2014	Height	MFCC	TIMIT
[29]	Shivakumar et al., 2014	Age	i-vectors	NIST SRE
[8]	Hansen et al., 2015	Height	MFCC + LSF + Formants	TIMIT
[30]	Poorjam, Bahari, & Vasilakakis, 2015	Height	i-vectors (MFCC)	NIST SRE
[31]	Fedorova et al., 2015	Age	i-vectors (MFCC) + i-vectors (SDCC)	NIST SRE
[32]	Galgali et al., 2015	Height, Age, Gender, and weight.	MFCC	Private dataset
[33]	Grzybowska & Kacprzak, 2016	Age	i-vectors (MFCC) + openSMILE	aGender
[34]	Sadjadi et al., 2016	Age	i-vectors	NIST SRE
[15]	Kalluri et al., 2016	Height, weight, shoulder width, waist size	MFCC	AFDS
[35]	Babu & Vijayasenan, 2017	Height, weight, shoulder size, waist size	GMM-supervectors(MFCC + Delta MFCC + Delta–Delta MFCC)	AFDS
[36]	Ghahremani et al., 2018	Age	x-vectors	NIST SRE
[37]	Zazo et al., 2018	Age	MFCC +Pitch + NCCF + POV	NIST SRE
[38]	Mallouh et al., 2018	Age and Gender	T-MFCC	aGender
[3]	(Beke, 2018	Height, Age, and Gender	F0 + MFCC+ Frequency modulation + Spectrum statistical features	BEA

[39]	Sánchez-Hevia et al., 2019	Age and Gender	MFCC	Common voice
[40]	Kalluri et al., 2019	Height and Age, weight, shoulder width, waist size	GMM-supervectors(MFCC + Delta MFCC + Delta–Delta MFCC)	TIMIT
[41]	Markitantov, 2020	Age and Gender	MFCC + Mel spectrogram	aGender
[11]	Kalluri et al., 2020	Height and Age	MFCC + Delta MFCC + Delta–Delta MFCC+ formants + harmonic features	TIMIT
[42]	Kwasny & Hemmerling, 2020	Age and Gender	x-vector	TIMIT
[43]	Kaushik et al., 2021	Height, Age	Filter banks + Pitch	TIMIT
[44]	Kwasny & Hemmerling, 2021	Age and Gender	x-vectors	Several datasets
[45]	A. Badr & K. Abdul-Hassan, 2021a	Age	MFCC + Delta MFCC + Delta–Delta MFCC +LPC + Delta LPC + Delta–Delta LPC + SCC+ Formants	Voxceleb1
[46]	A. Badr & K. Abdul-Hassan, 2021b	Age	MFCC + Delta MFCC + Delta–Delta MFCC +LPC + Delta LPC + Delta–Delta LPC + SCC+ Formants	TIMIT
[47]	Rajaa et al., 2021	Height, Age, and Gender	Wav2vec embeddings	NISP

4.1 Linear Predictive Coefficients LPC/ LPCC

LPC is a speech analysis method that provides a mathematical model to represent the human voice production system. In this model, the current speech sample can be closely estimated as a linear combination of previous samples using a set of coefficients that are determined by minimizing the error between the actual speech samples and the predicted ones. These coefficients represent a linear filter that models the vocal tract. Transforming LPC coefficients to cepstral coefficients produces LPCC. These features accurately approximate speech spectra, formants, and pitch [48].

4.2 Mel Frequency Cepstral Coefficients MFCC

MFCC is one of the key features employed in ASP research and is extensively used in different speaker profiling tasks. Either as a standalone feature or in combination with other features.

These features are obtained by first framing and windowing the speech signal into (20-30) ms frames, and then applying discrete Fourier transform (DFT) to each frame. To calculate the magnitude spectrum, these frames are then passed through a Mel-filter bank, resulting in a Mel spectrogram. Log of the Mel spectrogram is then measured and transformed into cepstral coefficients with a discrete cosine transform (DCT)[23].

As they only contain information from a given frame, MFCC features are considered static features. Dynamic features, on the other hand, are obtained by calculating the first and second derivatives to represent the temporal changes in the signal. The first-order derivative is called delta coefficients, and the second-order derivative is called delta-delta coefficients [48, 49]. MFCC's importance to different tasks of speaker profiling has been assessed in many studies. For example in [50], a study of audio feature selection, showed the high relevance of not only MFCCs but statistical functional parameters derived from the MFCCs to the task of height estimation. Furthermore, [24] studied the correlation of a speaker's height and vocal track length with different sets of features, and found that 57.15% of the speaker's height variability can be attributed to MFCC combined with other features such as formant frequencies.

In a similar effort,[7] compared the performance of MFCC and Linear predictive Coding coefficients (LPCC) and found that MFCC outperformed LPCC for age classification. Similarly, [27] compared the correlation of different features including (MFCC and LPCC) to height estimation and found that MFCC outperformed LPC.

In a study by [32], 26 MFCC coefficients were extracted and assessed in a multi-task setting to estimate height, weight, and age. The analysis revealed that only the first 13 coefficients are required for the task. The study also revealed the negative correlation between MFCC features and weight, while the correlation of height and age to MFCC is not clear.

In an interesting approach to obtaining MFCC features, [38] trained a deep Bottleneck feature extractor (DBF) to obtain T-MFCC features. When compared to the original MFCC features, the new T-MFCC features showed an increase in accuracy in age and gender classification.

4.3 Formant Frequencies

Formants are peaks in the spectrum with high acoustic energy around specific frequencies that corresponds to resonance in the human vocal tract [51]. The importance of formants in speaker profiling comes from the assumption that formant frequencies are inversely related to the length of the vocal tract [49, 52].

Formants have been closely related to phonic information in speech, as formants are phone-dependent, and different vowels produce different formants. Nonetheless, not all formants are equally relevant to height estimation. Out of the four first formants, F3 and F4 showed a higher correlation to height estimation than other formants [53], while in another study, F1 was found to have the highest correlation to a speaker's height [24]. This variation is attributed to the level of accuracy in formant measurement.

Other studies demonstrated the high correlation between specific vowels and height estimation [8, 27].

In multi-task learning of height, age, and other physical parameters [11], the first four formants are extracted and the correlation of the differences between formants and the fundamental frequency is assessed. It is reported that the difference between the fundamental frequency and f2, and f4 have a weak correlation for male and female speakers with height values. Log formant frequencies are utilized along with other features and showed good performance of 5.2 MAE, 6.7 RMSE for height estimation, and 5.4 MAE and 8.5 RMSE for age estimation when combined with harmonic features and fundamental frequency statistics.

4.4 Fundamental Frequency / Pitch

Pitch is the perceptual feature of the fundamental frequency (F0), and the terms are used interchangeably in the literature [54]. The fundamental frequency (F0), measured in Hertz, is defined as the number of cycles of opening and closing the glottis during a given period [55]. Estimation of the fundamental frequency, otherwise known as pitch detection algorithms, can be classified into three main categories: time domain, spectral domain, and hybrid. The time-domain autocorrelation method is the most common method of use [56]. The autocorrelation approach computes the dot-product of the speech signal with a shifted version of that signal [57]. Studies of F0 in speaker profiling show F0 is inversely proportional to the height of a speaker [58] [53].

The range of values of the f0 is affected by several parameters, such as gender, age, and body size [3]. Male speakers usually have lower frequencies than female speakers, and the f0 tends to decrease with age [9], especially for female speakers [10]. However, several studies demonstrated that even though the F0 itself is not significant in height estimations, statistics derived from the f0 as mean, and standard deviation are important to the task.[1, 11].

4.5 Voice quality features

Voice quality features are the distinct components of the voice that distinguish one speaker's voice from another's. These include the Harmonic to Noise Ratio (HNR), which measures the ratio between periodic and non-periodic components of a speech sound, speech rate (which represents syllables per second), the number of pauses, and duration of pause relative to the utterance length. Additionally, jitter and shimmer are two features that represent variations that occur in fundamental frequency, where Jitter refers to perturbation of the f0 while shimmer relates to random variations in amplitude or the intensity of the sound [59].

4.6 Gaussian Mixture Model supervectors

The Gaussian Mixture Model (GMM) is an unsupervised machine learning model that produces a finite number of multivariate gaussian components. A GMM is trained with an expectation-maximization algorithm (EM) that creates gaussian mixtures by updating gaussian means based on the maximum likelihood estimate.

A Universal Background Model (UBM) is a high-order GMM with a large number of gaussian mixtures, trained on a large set of speakers to learn the speaker-independent distribution of features.

In ASP every speech utterance is modeled with a GMM feature vector denoted as a supervector, where each speech utterance is represented by a weighted mixture of K Gaussian components, parameterized by mean vectors and a diagonal covariance matrix. A supervector is the result of concatenating the mean of each mixture [35].

As shown in Figure 7, to obtain supervector features, a UBM is trained on feature vectors such as MFCC or Mel spectrogram features. Then, an adaptation model for generating a speaker-dependant model is performed for each speech utterance. Given a feature vector X=[x_1,x_2,x_3,…,x_T], with T frames, the GMM probability density function is given as:

$$p_{UBM}(x) = \sum_{k=1}^{K} w_k N(x|\mu_{k,C_k}\ ) \qquad (3)$$

where N is a Gaussian, with weight wk, parameterized by the mean $\boldsymbol{\mu}_k$ and the covariance matrix C_k of the kth component respectively.

for any frame x_i of input feature vector X, the mean of all GMM components is calculated, and the resulting supervector is the concatenation of these means that represent the speech utterance.

For an input feature vector with F frames, and K GMM components, the supervector is the result of concatenating an F-dimensional GMM mean vector.

The GMM supervectors can be computed with frame-level MFCCs, or Mel spectrograms [25]. Alternatively, the supervectors can be extracted from the weighted Hidden Markov model (HMM) [26]. Additionally, three different types of supervectors: UBM weight posterior probability supervectors, GMM maximum likelihood linear regression (MLLR) matrix supervectors, and GMM mean supervectors, are employed for age and gender estimation in [60].

GMMs are considered computationally intensive, because of the frame-level calculations of the feature vector. The resulting supervectors also suffer from high dimensionality, as the adaptation process account for speaker-specific attributes in addition to session and channel variabilities.

In MTL, first-order statistics of a GMM-UBM are employed to learn the physical attributes of the speaker [35]. Dimensionality reduction is applied to the feature set since UBM-GMM produces feature vectors of high dimensionality, and the authors report an improvement in performance over their earlier work with bag of word (BoW) approach [15]. The authors also report a degradation in performance based on the spoken language used for training This implies that the performance of GMM models may be affected by phonetically different speech utterances.

An alternative approach to utilizing supervector features, is by initializing the weights of a deep neural Network (DNN) with a support vector regression (SVR) trained with GMM-UBM features [40].

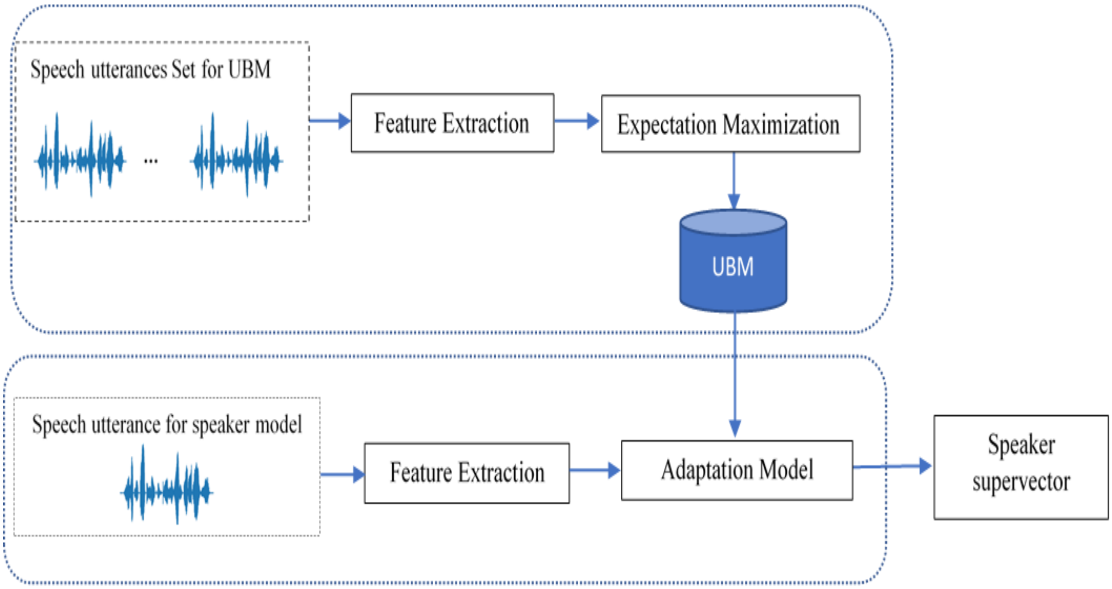

Figure 7: Steps involved in Supervector feature extraction

4.7 I-vectors

To address the shortcomings of the aforementioned supervectors, Dehak et.al.[61] proposed a low-dimensional feature vector that factors channel information out called the total factor vector, also referred to as i-vectors.

As mentioned earlier, a supervector M contains speaker-dependant and channel-dependant information that can be defined as:

$$M = u + Tv \qquad (4)$$

Where M is the speaker and channel independent supervectors from the UBM, T is rectangular, low dimensional (total variability matrix), and v is the total factors - intermediate vector or i-vector.

I-vectors, represent a spectral signature for a particular speech utterance, based on frame-level feature distribution in a low-dimensional form. Although efficient and widely used for speaker profiling, i-vectors suffer from the disadvantage of being less effective with short utterances.

In Speaker profiling, i-vectors are mostly employed for age prediction and classification, Some studies suggest that the use of i-vectors in combination with other acoustic features for both age estimation and age classification presented better results than using i-vectors alone [33]. Alternatively, an emerging trend in speaker profiling investigates the use of Deep Learning methods for feature extraction and speaker representation. Deep neural network-based i-vector modelling, are employed as an alternative to the common approach of GMM [34]. This approach produced phonetically-aware i-vectors that are then fed to an SVR model for age estimation.

Poorjam et.al. in [2, 14, 30] explored the use of i-vectors in multitasking speaker profiling for age, height, and weight estimation, and smoking detection. Each utterance was represented by an i-vector extracted from GMM weights and Non-negative Factor Analysis (NFA) for GMM weights adaptation and decomposition.

4.8 x-vectors

The x-vector architecture is DNN embedding that converts a variable-length speech utterance into a fixed-length embedding that contains speaker representation. This embedding is extracted using a time-delay NN (TDNN), which is a multi-layer NN that accounts for temporal context in the speech utterance by using frames of different time shifts at each layer and then summarizing the information using a statistics pooling layer [62].

As described in [63] and presented in Figure 8, the model consists of three main blocks, Frame-level, segment-level, and utterance-level, with an input layer that receives any acoustic feature vector such as MFCC, LPCC, etc.

The frame-level block consists of several layers with varying time shifts that act as frame-level feature extractors. The segment-level part starts with a statistical pooling layer that calculates the mean and the standard deviation of the output from the frame-level block, stack them together, and passes them as input to the next layers that constitute the x-vectors. the last block is utterance-level that gives the final decision output.

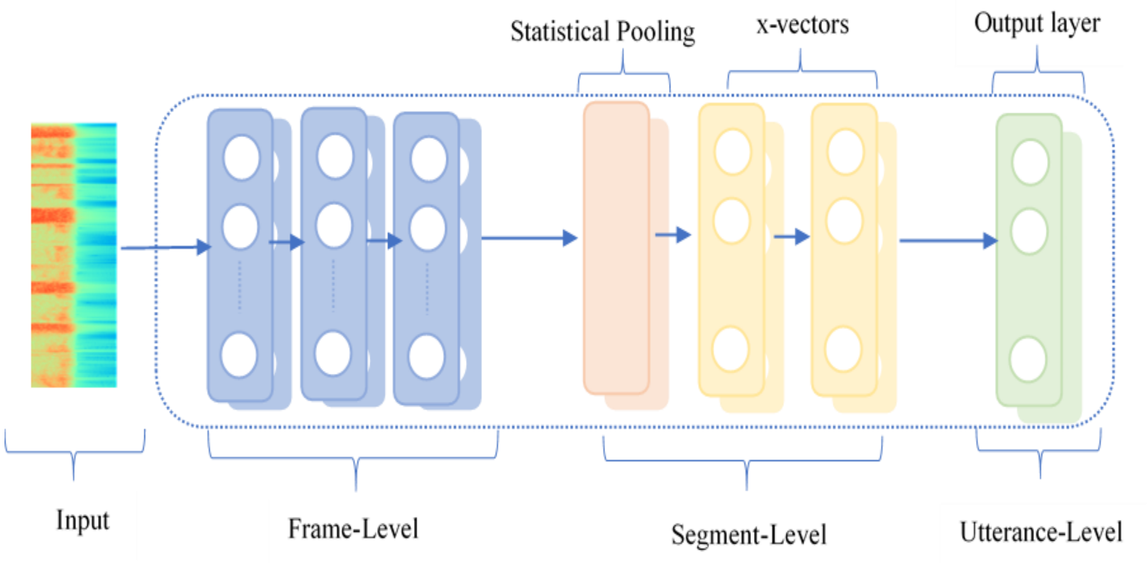

Figure 8: overview of the TDNN architecture and x-vectors

X-vectors have been employed in speaker profiling for age and gender estimation and produced promising outcomes [36, 42, 44]. DNN as front-end processing compared to i-vectors baseline and a fusion between x-vectors and i-vectors, where both are concatenated and processed by the DNN model The x-vectors approach outperformed both i-vectors and the fusion. However, the DNN approaches are data greedy and require longer training and processing time.

5. Learning Models

Learning models employed in ASP can vary between classification and regression models. A classification model assigns a label to the speaker from a predefined set of labels, for instance, in gender classification, the labels are male or female. For age classification the possible labels can be (young, adult, senior, etc.) and the same goes for height and weight. The assigned labels can come directly from the dataset used, or established based on some distribution of the data[8, 27].

On the other hand, regression models predict a numerical value for the estimation of exact age, height, and weight instead of assigning a label. In either case, the model construction consists of two phases: training and testing. In the training phase, feature extraction and selection are performed on speech utterances from the training set. The resulting feature vectors are used to train and fine-tune the model. In the test phase, the same procedure of constructing the feature vector is repeated on the test data and supplied to the model for classification or regression. The resulting outcome is compared against ground truth data and the model is evaluated using some evaluation metric.

The most common evaluation metrics used in ASP for regression tasks are mean absolute error (MAE), Root means squared error (RMSE). For classification, accuracies, and unweighted accuracies (UA) are used to evaluate classification models. Table 4 describes the evaluation metrics, and Table 5 describes the most used learning models in ASP literature.

Table 4: Evaluation Metrics in ASP Research

Metric	Formula	Description
Mean Absolute Error (MAE)	$\frac{\sum \lvert y_i - x_i \rvert}{n}$	Where: y_i is the predicted value, x_i is the target value, and n is the number of observations
Root mean square error (RMSE)	$\sqrt{\frac{\sum (y_i - x_i)^2}{n}}$	Where: y_i is the predicted value, x_i is the target value, and n is the number of observations
Accuracy	$\frac{TP + TN}{TN + TP + FN + FP}$	Measure the number of correctly classified instances by a particular classification. Where TP is the number of true positives, TN is the number of True negatives, FN is the number of false negatives, and FP is the number of false negatives.

Table 5: Machine Learning Algorithms employed in ASP

Model	Description	Used in
SVM	Support Vector Machine (SVM) is a supervised Machine learning algorithm for classification. SVM can handle linear and non-linear data by transforming the data into a higher dimension using a kernel function and finding a hyperplane that best separates the data points. The data points that are closest to the hyperplane are referred to as support vectors.	[7],[32, 57], [60]
SVR	Support vector regression is a variant of SVR that performs regression on the data (i.e., predicts a continuous value). While regression algorithms attempt to minimize errors in training, SVR aims to fit the error within a specific threshold.	[1],[50], [11, 15, 35], [34]
LSSVR	Least square support vector regression (LSSVR) is a variant of SVR that is also used for regression problems. LSSRVR is preferred for its faster training process due to solving a linear equation system instead of the quadratic programming problem of SVR. Additionally, the LSSVR requires fewer tuning parameters. Although solving linear equations is faster and requires fewer computations and memory than that solving quadratic programming optimization problems, the simplicity of LSSVR comes with the cost of lack of sparsity, which means that the model can get unnecessarily large	[25], [30], [2], [33]
GMM	A Gaussian mixture model (GMM) can be thought of as a generalization of the k-means clustering algorithm. It is a probabilistic model that assumes the dataset to consist of a mixture of several Gaussian distributions.	[6],[27],[28],[8],[38]
ANN	ANN is a broad term that includes several concepts such as MLP (multilayer perceptron) and DNN (Deep Neural Network). MLP usually consists of one input layer, one output layer, and several hidden layers where each layer consists of several neurons. ANN is known for its capability to learn complex nonlinear functions from input data. Although ANNs are robust and efficient, the effectiveness of an ANN is affected by several parameters, to mention a few, the number of layers, the number of neurons in a layer, learning rate, and the training algorithm. Moreover, ANN typically requires a large amount of training data to achieve the required performance.	[31],[2]
LR	Linear regression is a linear model used to predict the value of a variable based on the value of another variable.	[24], [1],[8, 27],[32],[3]

DT	Decision trees (DT) are a family of algorithms that perform classification and regression by deducing decision rules from the data. DTs employed in ASP include C4.5 and M5.	[7],[1]
Ensemble	Ensemble learning is a machine learning approach that employs multiple learning models and incorporates their outcomes to improve performance.	[1],[3]

6. Discussion

A careful review of ASP literature reveals some insights into the process of feature extraction and learning from speech utterances concerning different tasks of speaker profiling. Several factors affect the performance of an ASP system, some of these factors are discussed in this section.

Effects of Dataset Choice:

The quality of data used to train any model represents an important factor in the performance of the trained model. Most of the datasets employed by speaker profiling systems reviewed in this work are not designed for this task, thus there are many limitations to these data sets in terms of data imbalance or the insufficient amount of data for training. NISP is the only dataset designed specifically for ASP; however, the dataset is relatively new, and few studies employ the dataset for evaluation. As results obtained from different ASP systems are highly dependent on the dataset used in the training process, the range of these results varies significantly, rendering these studies incomparable.

A number of approaches have been proposed to address the problem of limited training data.. For instance, [42, 44] employed x-vectors and d-vectors in the joint estimation of age and gender, combined with transfer learning from networks pre-trained for different tasks other than speaker profiling, utilizing well-known speaker recognition datasets. The use of transfer learning showed an improvement in terms of MAE and RMSE.

Similarly, to address data limitations, [40] explored the use of DNN initialization with weights obtained from an SVR model trained with GMM mean supervectors for joint height and age estimation.

Other works that explored the use of transfer learning in joint age and gender classification [41, 42, 44], exploited pre-trained image models to classify speakers based on images of MFCC and Mel spectrogram features. However, the authors found that a TDNN showed better accuracy than those models.

Transfer Learning has also been utilized to learn speaker representations as speech encoders in a semi-supervised manner using speaker recognition datasets such as librispeech, by using LSTMs in conjunction with CNN [47]. The encoded speech is then fed into a two-layer NN for fine-tuning and joint estimation of height age, and gender.

Effects of Feature Representation Approach:

The performance of the ASP system is mainly dependent on the features extracted from speech signals. As can be seen in Figure. 9, a trend can be observed in the types of features selected for different tasks of speaker profiling. For age estimation, the features are mostly i-vectors and supervectors, whereas, for height estimation, handcrafted features are mostly used. However, MFCC features are the dominant feature representation selected for most tasks of

ASP. MTL learning, on the other hand, tries to combine these different representations. Some studies found that individual features perform similarly and achieve comparable results, while the combination of these features shows significant improvement, and these features are found to be complementary to each other [11, 33].

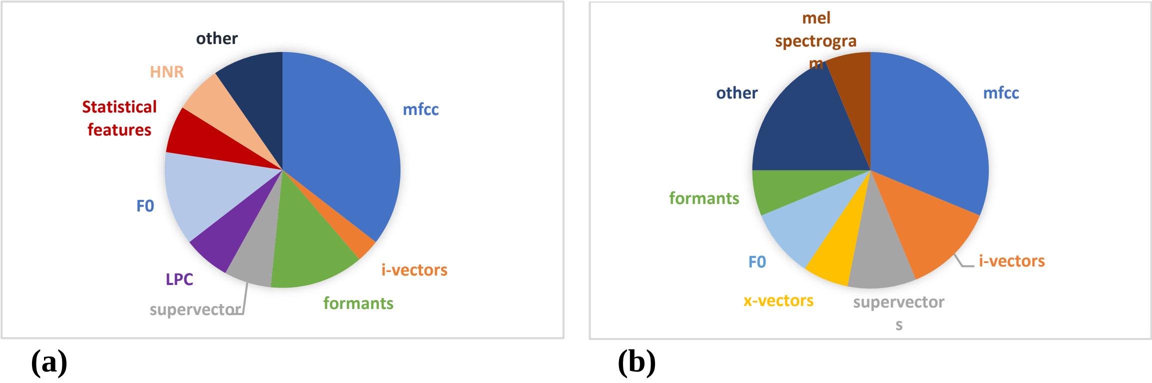

Figure 9: Distribution of features utilized for (a) Height estimation, (b) Age estimation in reviewed articles

In addition to traditional feature extraction methods, phone-based features have also been employed in ASP with promising prospects. Several studies [24] [8, 27] explored the effect of different vowels on height estimation and found specific vowels that showed a high correlation to a speaker's height.

Similarly, A phonetically-aware i-vector was proposed by [34] for age estimation using a DNN acoustic model based on context-dependent triphones to extract i-vectors as opposed to a GMM model.

Moreover, [43] investigated the use of DNN for height and age estimation by proposing an LSTM architecture with an attention mechanism. The authors note that vowel information played a main role in the estimation accuracy.

Phone-based approaches' effectiveness in ASP comes with the disadvantage of relying on specific vowels and may require speech transcription, which in turn may not be practical in some applications. Additionally, they may not be effective for continuous speech, where they require the speaker to utter isolated vowels.

Another factor that affects the performance of ASP is the high dimensionality of features. Dimensionality reduction (such as principle component analysis and LDA), and feature selection algorithms employed in several studies, such as [35, 46, 57], reported an increase in performance and reduction in computation time.

The current research is limited in the types of learning models and features used for ASP as compared to other fields of speech processing such as speech recognition and speaker identification, and different feature extraction and selection methods are yet to be explored.

Effects of Learning Model Choice

Several studies evaluated the effect of the learning model on ASP performance. While some studies reported model choice didn't affect the performance [34], other studies found significant improvements when combining different models. For instance, Spectral and temporal features are employed by [3] with different learning models, including (Linear regression, random forests, GBM, SVR, and DNN). The worst performing model was linear regression, while the best results was obtained from ensemble learning trained on the results of all the five learning models. In a similar effort to examine the effect of model choice on height estimation, [1] examined different regression algorithms including linear regression, Support vector Regression, Multi-Layer Perceptron, AR, M5, and Bagging where each speech utterance is transformed into a feature vector with different statistical features using openSMILE toolkit. Out of all the tested models, Bagging was the best performing model.

Similarly, the fusion of regression and classification is found to improve the performance of different ASP tasks as seen in [8, 27, 64]. For instance, The authors in [33] report an improvement in the fusion of age regression and age classification systems. The fusion is performed by calculating the cosine distance on the output from the regression system after mapping the result to one of the age classes defined for the classification system.

The study in [27] proposed a fusion between a classification approach that assigns a height class for a speaker with a confidence score, and linear regression approach with smoothed formant tracks as features. Experimental results showed that the fusion of the two systems outperforms the results obtained from the regression system alone, indicating the complementary effect of the regression and classification approaches.

Other factors may affect the performance of the model, such as single-task learning or multi-task learning. Several studies reported a relative improvement in a multi-task setting over single task estimation [2, 43, 47].

7. Challenges and Future directions

This review identified several challenges that face the research on ASP. The advancement of ASP requires collaborative research effort in addressing these challenges and enhancing the performance of speaker profiling systems. These research challenges are discussed below:

The need for generating comprehensive and high quality datasets for ASP:

One of the main challenges that hinder ASP research is the availability of appropriate datasets with comprehensive information about speakers. Most research in ASP relies on existing databases with limited data about speakers, which prevents true multitask learning of several features at once. Moreover, different studies report their results according to the quality of data used, making the comparison of the results of different studies unfeasible. Thus, benchmark datasets are needed to evaluate the performance of different approaches. Additionally, modern deep learning approaches require massive amounts of speech data to reach a satisfactory performance in ASP and arriving at a representative feature set for different traits of the speaker. Data augmentation can address this limitation by transforming existing samples and creating new ones. Another approach is the use of transfer learning and semi-supervised learning, where models are pre-trained on various speech datasets.

The need for focusing on real-time performance in ASP:

ASP applications require a timely response, whether it was designed for real-time applications such as customer service, call routing, human-robot interaction, or forensic applications. Most research in ASP target high accuracy without regard to speed. Simpler modelling techniques that can operate in real-time with a compressed form of speech utterance are highly required to accommodate the requirements of ASP applications.

The need for DNN architectures that suites ASP:

Results obtained with DNN approaches proved effective and produce good results compared to previous studies. However, most studies adopt architectures from other fields, such as image processing, to handle ASP speech data. There is an urgent need to develop guidelines or standards on the appropriate architecture for speech data analysis.

8. Conclusion

Automatic speaker profiling is an emerging research domain in speech processing that offers promising prospects in various real-life applications. Therefore, this work provided an insight into the field of automatic speaker profiling with its subfields of gender detection, age, height, weight estimation, and other parameters. This paper explores various aspects of ASP research, such as datasets available for ASP, the most common features used in ASP, the learning models applied to it, and the challenges facing research in the field.

One of the key elements of any speaker profiling task is the speech dataset used and the metadata contained in that dataset. Datasets that can be employed in different tasks of ASP are presented with their parameters and shortcomings.

Another key factor of ASP is feature extraction and representation, where the performance of ASP is heavily dependent on the type of extracted features. The different extraction methods and speech representation methods are discussed. Furthermore, different tasks of ASP research are analysed based on the different classifiers and regression models used in learning the speaker profile from features. The performance of these studies is then analysed and discussed. Finally, the paper presents the challenges and limitations that face ASP research.

References

[1] I. Mporas and T. Ganchev, "Estimation of unknown speaker’s height from speech," *International Journal of Speech Technology,* vol. 12, no. 4, pp. 149-160, 2010, doi: 10.1007/s10772-010-9064-2.

[2] A. H. Poorjam and M. H. Bahari, "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)*, 2014: IEEE, pp. 7-12.

[3] A. Beke, "Forensic speaker profiling in a Hungarian speech corpus," in *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2018: IEEE, pp. 000379-000384.

[4] C. Müller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," in *Ninth International Conference on Spoken Language Processing*, 2006: Interspeech, pp. 2118–2121

[5] C. Müller and F. Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," in *Eighth Annual Conference of the International Speech Communication Association*, 2007: INTERSPEECH pp. 2277–2280.

[6] H.-J. Kim, K. Bae, and H.-S. Yoon, "Age and gender classification for a home-robot service," in *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 2007: IEEE, pp. 122-126.

[7] M.-W. Lee and K.-C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 3, no. 12, 2012.

[8] J. H. Hansen, K. Williams, and H. Boril, "Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models," *J Acoust Soc Am,* vol. 138, no. 2, pp. 1052-67, Aug 2015, doi: 10.1121/1.4927554.

[9] M. Nishio and S. Niimi, "Changes in speaking fundamental frequency characteristics with aging," *Folia phoniatrica et logopaedica,* vol. 60, no. 3, pp. 120-127, 2008.

[10] J. T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian, "Effects of aging on vocal fundamental frequency and vowel formants in men and women," *Journal of Voice*, vol. 32, no. 5, pp. 644. e1-644. e9, 2018.

[11] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," *Speech Communication*, vol. 121, pp. 16-28, 2020, doi: 10.1016/j.specom.2020.03.008.

[12] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: a study using magnetic resonance imaging," *J Acoust Soc Am*, vol. 106, no. 3 Pt 1, pp. 1511-22, Sep 1999, doi: 10.1121/1.427148.

[13] J. Laver and P. Trudgill, "Phonetic and linguistic markers in speech," *Social markers in speech*, vol. 1, p. 32, 1979.

[14] A. H. Poorjam, M. H. Bahari, and H. Van Hamme, "Speaker weight estimation from speech signals using a fusion of the i-vector and NFA frameworks," in *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, 2015: IEEE, pp. 118-123.

[15] S. B. Kalluri, A. Vijayakumar, D. Vijayasenan, and R. Singh, "Estimating multiple physical parameters from speech data," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016: IEEE, pp. 1-5.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

[17] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective," *Speech communication*, vol. 31, no. 2-3, pp. 225-254, 2000.

[18] S. B. Kalluri, D. Vijayasenan, S. Ganapathy, and P. Krishnan, "NISP: A Multi-lingual Multi-accent Dataset for Speaker Profiling," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 6953-6957.

[19] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, pp. 1562–1565.

[20] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[22] H. Fayek, "Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what’s in-between," *URL: https://haythamfayek. com/2016/04/21/speech-processingfor-machine-learning. html*, 2016.

[23] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114591, 2021.

[24] S. Dusan, "Estimation of speaker's height and vocal tract length from speech signal," in *Ninth European Conference on Speech Communication and Technology*, 2005, pp. 1989–1992.

[25] M. H. Bahari and 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: IEEE, pp. 1-6.

[26] M. H. Bahari, "Speaker age estimation using Hidden Markov Model weight supervectors," in *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, 2012: IEEE, pp. 517-521.

[27] K. A. Williams and J. H. Hansen, "Speaker height estimation combining GMM and linear regression subsystems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013: IEEE, pp. 7552-7556.

[28] H. Arsikere, S. M. Lulich, and A. Alwan, "Estimating Speaker Height and Subglottal Resonances Using MFCCs and GMMs," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 159-162, 2014, doi: 10.1109/lsp.2013.2295397.

[29] P. G. Shivakumar, M. Li, V. Dhandhania, and S. S. Narayanan, "Simplified and supervised i-vector modeling for speaker age regression," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014: IEEE, pp. 4833-4837.

[30] A. H. Poorjam, M. H. Bahari, and V. Vasilakakis, "Height estimation from speech signals using i-vectors and least-squares support vector regression," in *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, 2015: IEEE, pp. 1-5.

[31] A. Fedorova, O. Glembek, T. Kinnunen, and P. Matějka, "Exploring ANN back-ends for i-vector based speaker age estimation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 3036–3040.

[32] S. Galgali, S. S. Priyanka, B. Shashank, and 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)*, 2015: IEEE, pp. 486-489.

[33] J. Grzybowska and S. Kacprzak, "Speaker Age Classification and Regression Using i-Vectors," presented at the Interspeech 2016, 2016.

[34] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: IEEE, pp. 5040-5044.

[35] K. S. Babu and D. Vijayasenan, "Robust features for automatic estimation of physical parameters from speech," in *TENCON 2017-2017 IEEE Region 10 Conference*, 2017: IEEE, pp. 1515-1519.

[36] P. Ghahremani *et al.*, "End-to-end Deep Neural Network Age Estimation," presented at the Interspeech 2018, 2018.

[37] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks," *IEEE Access*, vol. 6, pp. 22524-22530, 2018, doi: 10.1109/access.2018.2816163.

[38] A. A. Mallouh, Z. Qawaqneh, and 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*, vol. 30, no. 8, pp. 2581-2593, 2018, doi: 10.1007/s00521-017-2848-4.

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

[40] S. B. Kalluri, D. Vijayasenan, and 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: IEEE, pp. 6580-6584.

[41] M. Markitantov, "Transfer learning in speaker's age and gender recognition," in *International Conference on Speech and Computer*, 2020: Springer, pp. 326-335.

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

[43] M. Kaushik, V. T. Pham, and E. S. Chng, "End-to-end speaker height and age estimation using attention mechanism with LSTM-RNN," *arXiv preprint arXiv:2101.05056*, 2021.

[44] D. Kwasny and D. Hemmerling, "Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks," *Sensors (Basel)*, vol. 21, no. 14, Jul 13 2021, doi: 10.3390/s21144785.

[45] A. A. Badr and A. K. Abdul-Hassan, "Age Estimation in Short Speech Utterances Based on Bidirectional Gated-Recurrent Neural Networks," *Engineering and Technology Journal*, vol. 39, no. 1B, pp. 129-140, 2021, doi: 10.30684/etj.v39i1B.1905.

[46] A. A. Badr and A. K. Abdul-Hassan, "Estimating Age in Short Utterances Based on Multi-Class Classification Approach," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 1713-1729, 2021, doi: 10.32604/cmc.2021.016732.

[47] S. Rajaa, P. Van Tung, and C. E. Siong, "Learning Speaker Representation with Semi-supervised Learning approach for Speaker Profiling," *arXiv preprint arXiv:2110.13653*, 2021.

[48] R. Chaudhary, "Short-term spectral feature extraction and their fusion in text independent speaker recognition: A review," *BVICA M's International Journal of Information Technology*, vol. 5, no. 2, p. 630, 2013.

[49] T. Kinnunen, "Spectral features for automatic text-independent speaker recognition," *Licentiate's thesis*, 2003.

[50] T. Ganchev, I. Mporas, and N. Fakotakis, "Audio features selection for automatic height estimation from speech," in *Hellenic Conference on Artificial Intelligence*, 2010: Springer, pp. 81-90.

[51] P. A. Abhang, B. W. Gawali, and S. C. Mehrotra, "Technical aspects of brain rhythms and speech parameters," *Introduction to EEG-and speech-based emotion recognition*, pp. 51-79, 2016.

[52] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996, vol. 1: IEEE, pp. 353-356.

[53] R. Greisbach, "Estimation of speaker height from formant frequencies," *International Journal of Speech Language and the Law,* vol. 6, no. 2, pp. 265-277, 1999.

[54] D. Gerhard, "Pitch extraction and fundamental frequency: History and current techniques," in " " Department of Computer Science, University of Regina Regina, SK, Canada, Regina, SK, TR-CS 2003-06, 2003.

[55] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis–jitter, shimmer and hnr parameters," *Procedia Technology,* vol. 9, pp. 1112-1122, 2013.

[56] M. A. R. Hasan, R. Yasmin, D. Das, M. Hoque, M. Pramanik, and M. Rahman, "Fundamental Frequency Extraction of Noisy Speech Signals," *Rajshahi University Journal of Science and Engineering,* vol. 43, pp. 51-61, 2015.

[57] A. Badr and A. Abdul-Hassan, "CatBoost Machine Learning Based Feature Selection for Age and Gender Recognition in Short Speech Utterances," *International Journal of Intelligent Engineering and Systems,* vol. 14, no. 3, pp. 150-159, 2021, doi: 10.22266/ijies2021.0630.14.

[58] W. A. Van Dommelen and B. H. Moxness, "Acoustic parameters in speaker height and weight identification: sex-specific behaviour," *Language and speech,* vol. 38, no. 3, pp. 267-287, 1995.

[59] H. F. Wertzner, S. Schreiber, and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders," *Revista Brasileira de Otorrinolaringologia,* vol. 71, no. 5, pp. 582-588, 2005.

[60] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language,* vol. 27, no. 1, pp. 151-167, 2013, doi: 10.1016/j.csl.2012.01.008.

[61] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788-798, 2010.

[62] D. Sztahó, G. Szaszák, and A. Beke, "Deep learning methods in speaker recognition: a review," *arXiv preprint arXiv:1911.06615,* 2019.

[63] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP),* 2018: IEEE, pp. 5329-5333.

[64] O. Buyuk and M. L. Arslan, "Combination of Long-Term and Short-Term Features for Age Identification from Voice," *Advances in Electrical and Computer Engineering,* vol. 18, no. 2, pp. 101-108, 2018, doi: 10.4316/aece.2018.02013.