Received: 31 October 2020 | *Revised: 11 November 2020* | *Accepted: 12 November 2020*
DOI: 10.37917/ijeee.16.2.10 *Vol. 16| Issue 2| December 2020*

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Review Article

A Review on Voice-based Interface for Human-Robot Interaction

Ameer A. Badr*[1,2], Alia K. Abdul-Hassan[1]
[1] Department of Computer Science, University of Technology, Baghdad, Iraq
[2] College of Managerial and Financial Sciences, Imam Ja'afar Al-Sadiq University, Salahaddin, Iraq

Correspondence
*Ameer A. Badr
Department of Computer Science,
University of Technology, Baghdad, Iraq
Email: cs.19.53@grad.uotechnology.edu.iq

Abstract
With the recent developments of technology and the advances in artificial intelligence and machine learning techniques, it has become possible for the robot to understand and respond to voice as part of Human-Robot Interaction (HRI). The voice-based interface robot can recognize the speech information from humans so that it will be able to interact more naturally with its human counterpart in different environments. In this work, a review of the voice-based interface for HRI systems has been presented. The review focuses on voice-based perception in HRI systems from three facets, which are: feature extraction, dimensionality reduction, and semantic understanding. For feature extraction, numerous types of features have been reviewed in various domains, such as time, frequency, cepstral (i.e. implementing the inverse Fourier transform for the signal spectrum logarithm), and deep domains. For dimensionality reduction, subspace learning can be used to eliminate the redundancies of high-dimensional features by further processing extracted features to reflect their semantic information better. For semantic understanding, the aim is to infer from the extracted features the objects or human behaviors. Numerous types of semantic understanding have been reviewed, such as speech recognition, speaker recognition, speaker gender detection, speaker gender and age estimation, and speaker localization. Finally, some of the existing voice-based interface issues and recommendations for future works have been outlined.

KEYWORDS: Human-robot interaction (HRI), Social robotics, Voice-based perception, Voice-based semantic understanding, Voice-based features extraction.

I. INTRODUCTION

Social robotics, an effective robotics branch, has lately attracted increased interest in several disciplines such as mechatronics, computer vision, and artificial intelligence, whereas many social robots were evolved. A precise description of the social robot remains elusive and has been identified by various practitioners from different perspectives [1]. According to Yan et al. [1], one can define a social robot as follows: *"A social robot is a robot which can execute designated tasks, and the necessary condition turning a robot into a social robot is the ability to interact with humans by adhering to certain social cues and rules".*

Human-Robot Interaction (HRI) has currently received the attention of the research community, as well as of the industrial world. The HRI has been concerned with the lookup community as a subject of interaction and communication into human beings with robots. Robots have already started transferring out of the laboratory or manufactured environments into more complex human cause environments, such as homes, offices, hospitals, or even principal space. The layout, regarding appearance, robotic behavior, and cognitive, yet social advantage is tremendously challenging, yet requires interdisciplinary assistance between classic robotics, cognitive sciences, or psychology [2].

Perception methods can be considered as one of the most important capabilities in HRI, especially for a social robot which is an important branch of robotics. Of all the perception methods for robots, a voice-based perception has been of great importance recently. With such a growing number of applications involving social HRI, adaptive socially responsive speech interfaces are increasingly needed. In HRI, people prefer to view robots consciously as non-living mechanics but unconsciously engage robots in essentially social ways. People respond similarly when it comes to voice-based interfaces, by applying automatic and unconscious social responses [3].

The perception system of a social robot is to assist the robot realize the surroundings excellently. It can use audio signal, visual signal, laser reading signal, and tactile to interact with humans. Having gained these signals, it is helpful for robots' perception systems to know how to

process them to gain valuable information. Social robot's perception tasks require very important requirements, such as autonomy and low computational cost, besides high accuracy in recognition. For real-time applications, the methods of perception that have low computational costs are desirable, such that the robot can immediately respond to humans. The ultimate goal, for social robots, is to get them to work in real environments.

The main contributions of this work are highlighted and summarized as follows:

1) HRI taxonomies and the role of adaptivity in HRI have been described, in which one can distinguish between an autonomous robot and an adaptive robot.
2) The voice-based perception methods for HRI in social robots have been reviewed, which summarizes perception strategies out of three aspects: feature extraction, dimensionality reduction, and semantic understanding. Moreover, a review of voice-based feature and voice-based semantic understanding is presented.
3) Representative voice-based social robot methods and techniques have been reviewed.
4) The challenging issues in the voice-based interface for HRI have been addressed.

The organization of the remaining sections of this work is as follows. One can start by giving a brief definition of HRI and illustrate the HRI taxonomies that impact the interactions between humans and robots and the role of adaptivity in HRI in section two. Section three reviews the voice-based perception method from three aspects: feature extraction methods, dimensionality reduction methods, and semantic understanding methods. A review of representative voice-based social robots is presented in section four. Discussion and recommendations for future works are presented in section five. Finally, section six gives a brief conclusion of this work.

II. HUMAN-ROBOT INTERACTION (HRI)

HRI is an area of study devoted to designing, understanding, and evaluating robotic systems for utilization with or by humans. By definition, interaction means communication among humans and robots. This communication can be performed in many forms, but these forms are greatly affected by if the robot and the human whether they are in close proximity to each other or not. Thus, interaction or communication can be divided into two common categories [4]:

1) Proximate Interaction: Robots and humans are collocated.
2) Remote Interaction: The robots and the human are separated temporally or spatially.

By taking advantage of the categories above, one can distinguish between applications that need mobility, physical manipulation, or social interaction [4].

A. HRI Taxonomies

In essence, many taxonomies impact the interactions between humans and robots can be affected by a designer. Such taxonomies include task type, task criticality, robot morphology, a ratio of people to robots, composition of robot teams, level of shared interaction among teams, type of human-robot physical proximity, decision support for operators, autonomy level/amount of intervention, and information exchange [4], [5]. Of these taxonomies, the most used in HRI describes below:

1) *Autonomy Level / Amount of Intervention*: One of the factors determining the interaction between a robot and a human being is the amount of intervention required to control the robot, where the amount of intervention ranges from a remote control to complete self-control. If a robot has greater autonomy, hence, less interaction is required, while constant interaction is required at the teleoperation level. The percentage of time the robot spends on its task, measures the level of autonomy, while the percentage of time that the human operator controls a robot, and measures the level of intervention. These two measures sum to 100% [5].
2) *Interaction Roles:* The number of roles a person may play when interacting with robots is five: (supervisor, operator, teammate, programmer, and bystander). When a person's role is supervisory, he does not need direct control of the robot but only monitors its behavior, while the operator needs to interact more with the robot. Humans and robots can work in one team as a teammate to accomplish a certain task, while the programmer needs to modify or change robots' software or hardware. A bystander does not control the robot but only understands what the robot does [5].
3) *Type of Human-Robot Physical Proximity:* HRI may be at different distances depending on the tasks or purpose of this interaction. In the event that humans and the robot are collocated, there are five forms of physical proximity between the robot and humans (avoiding, passing, following, approaching, and touching), where these forms are arranged from least to most physical interaction [5].
4) *Robot Morphology:* In terms of the external appearance, robots can take many forms, as humans interact with the robot differently depending on their appearance. Robot morphology is given three values: anthropomorphic (i.e. The appearance here is somewhat similar to the human form), zoomorphic (i.e. The appearance here is related to the shape of an animal), and, Functional (i.e. The appearance here is related to the nature of the robot's function, neither human nor animal) [5].
5) *Information Exchange:* Here the efficiency of the HRI is measured by the time taken to communicate between the robot and the human being; the amount of information that is exchanged through this interaction, and the amount of common understanding or common ground between the robot and the human [4].

To determine the way information is exchanged between humans and robots, there are two main dimensions: the format of the communications and the communications medium. By three senses, namely hearing, touch, and vision

media was delineated. These media are presented in HRI as follows [4]:

1) *Gestures:* including facial and hand movements.
2) *Visual displays:* presented as graphical user interfaces or augmented reality interfaces.
3) *Natural language and speech:* which include both auditory speech and text-based responses.
4) *Physical interaction and haptics:* frequently used in teleoperation or in augmented reality to mention a sense of presence, and also frequently used proximately to promote social, emotional, and assistive exchanges.

The quality of the information that is exchanged between humans and robots may vary depending on domains. Communication can be through the natural language and speech based on a specific official language, or this language can be restricted to a sub-language or a specific domain. In Haptic and through this type of communication, warnings can be given by vibrations. Audio information presentation can include 3D awareness, auditory alerts, and a speech-based exchange of information. Graphical user interfaces display information in ways including windows-type interactions, ecological displays, and immersive virtual reality [4].

B. *Adaptivity in HRI*

To obtain an accurate definition of Adaptive Robot Interfaces (ARIs), it is necessary first to distinguish between an adaptive robot and an autonomous robot. ARI is an autonomous or semi-autonomous robot where the robot decision is made based on the awareness of environmental information from the user, where this information may include user profile, user emotions, user past interactions, and user personality. One or more of the following HRI adaptation capabilities could be used by ARI: Communication through dialogue, learning according to user responses, emotions understanding and exhibiting, establishing a social relationship, having different roles and social characteristics, and, reacting accordingly to different social situations [6].

According to Ahmad et al. [6], one of the biggest challenges in the field of social robotics is the development of an adaptive robotic system in the real world. Different applications of adaptive social robotic systems have been proposed by researchers over the last decade that was able to display one or more of the aforementioned abilities.

III. VOICE-BASED PERCEPTION IN HRI

One of the most important capabilities of a social robot is its perception. A social robot's HRI is composed of three parts: perception, action, and an intermediate mechanism. Perception represents the unit for acquiring and analyzing environmental information. An action represents the robot's response after it receives control signals. The intermediate mechanism linking perception and action to produce control signals is by using the results of the perception analysis module. Because it is considered a bridge between the social robot and the external environment, the perceptual system dominates an entire HRI. The better the robot understands the surrounding world; the more meaningful responses it can give [1]. This work reviews the voice-based perception method of HRI in social robots due to the great importance of a voice-based perceptual system.

As seen in Fig. 1, social robots' voice-based perception can be fundamentally included in three steps: feature extraction, dimensionality reduction, and semantic understanding. The feature extraction process aims to get feature descriptors for subsequent understanding tasks by converting a raw audio signal from the sensor. Dimensionality reduction aims to minimize the computation complexity after the feature extraction process. Semantic understanding aims to infer the behaviors of humans or objects from the extracted features. Standard semantic understanding tasks, for voice-based perception, include speech recognition, speaker recognition, speaker emotion recognition, rhythms recognition, speaker gender detection, speaker age estimation, and speaker localization. On the other hand, the social robot's action is represented by a speech response through the use of the Text to Speech (TTS) method [1].

A. *Feature Extraction*

One can define the feature extraction as the process of highlighting the most distinctive characteristics of a signal. The extracted feature is appropriate when it mimics the characteristics of the signal extracted from it, in a somewhat compact manner. The audio features evolution can be subclassified into time, frequency, joint time-frequency, and deep domains. Although the time domain features are among the oldest and simplest features, they still play a significant role in audio classification and analysis. Some frequency domain features have been developed to analyze the spectrum of audio signals and have been used in many applications so far. This was followed by developing the joint time-frequency feature extraction algorithms. On the other hand, the deep features are broadly used in various applications; deep features for the audio signal processing were used in the area of speaker recognition, audio-video analysis, and acoustic scene classification [7].

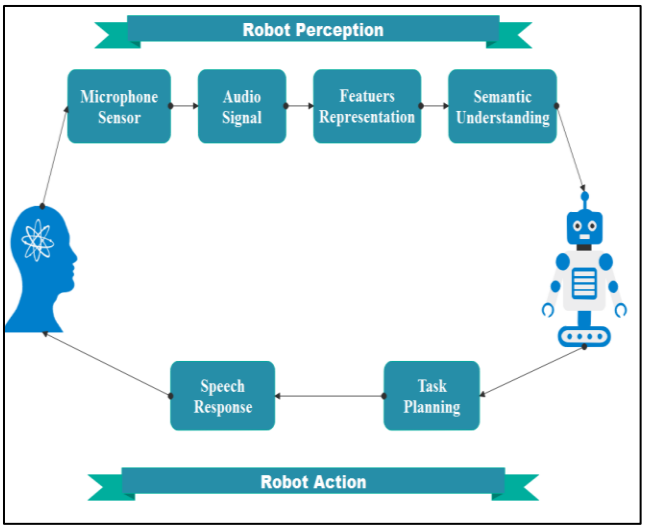

Fig. 1: Voice-based Perception in HRI.

The concept of windowing is very important when dealing with the time domain. All audio signals are considered as a time series signal which evolves with time; when the signal reflects stationary properties over time, the time domain analysis of this signal is simple. However, the audio signals are non-stationary in real-time; hence windowing technique must be employed to analyze such non-stationary signals when a long non-stationary signal is analyzed as short quasi-stationary signal chunks [7]. The most important time-domain features used in audio signals are shown in Fig. 2 and described below.

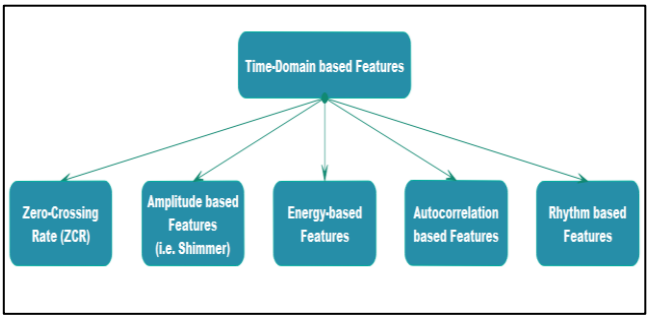

Fig. 2: Time-Domain based Audio Features.

1) *Zero-Crossing Rate (ZCR):* It is known as the rate of an audio frame's signal sign change. In other words, it is the number of times the signal changes from the negative to the positive, and vice versa, divided by the frame length. By using ZCR, one can determine if a speech frame is voiced, unvoiced, or silent because ZCR is higher in the unvoiced speech frames compared with voiced speech frames [7]. ZCR is used for various applications like a method for calculating speech signal Fundamental Frequency (F0), design discriminator and classifier, detects voice activity [8], [9], vowel analysis and detection [10], music/speech discrimination [11], and speaker gender and age recognition [12].
2) *Amplitude based Features:* This type of feature is based on signal temporal envelope analysis. The most important type of such feature is the Shimmer feature [7]. Shimmer calculates cycle-to-cycle amplitude variations within a waveform. It is used in speaker verification [13], speaker age classification [14], and speaker age and gender estimation [15], [16].
3) *Energy-based Features:* There are several types of an energy-based feature used in audio signals, the most used type of such features is Short Time Energy (STE) [7], [17]. As aforementioned, by using the framing and windowing method, one can transform non-stationary sound signals into small portions of quasi-stationary signals. STE can be defined as an average energy per frame. For unvoiced segments, STE is low, while for the voiced segment, STE is high. STE is used in speaker gender and age recognition [12], detect the unvoiced-voiced segments [9], vowel detection, and analysis [10].
4) *Autocorrelation based Features:* In the time domain, the measure of self-similarity between the signal and its delayed version can be defined as autocorrelation. The autocorrelation value of (+1), (-1), and (0) represents a positive association, a negative association, and no association respectively. It is widely used to estimate signal F0 [7].
5) *Rhythm-based:* Rhythm is generally a frequent recurrence of pattern over time. Rhythm can be found in musical instruments, environmental sounds, and speech [7]. This feature is sometimes used in HRI as in the dancing robot in [18].

In audio signal processing, one of the most important tools is the frequency domain analysis. As aforementioned, the time-domain shows the signal variation concerning time. Hence by using auto-regression or Fourier transform analysis, the time-domain signal is converted into a frequency-domain signal to analyze a signal in terms of frequency. The main frequency domain features are shown in Fig. 3 and discussed below [7], [19]:

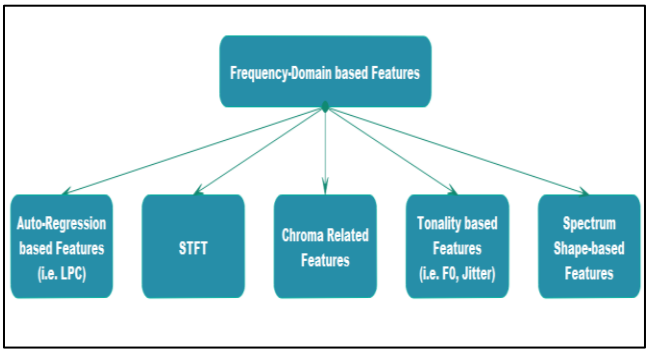

Fig. 3: Frequency-Domain based Audio Features.

1) *Auto-Regression based Features:* These types of features are extracted from linear prediction analysis of a signal. Linear Predictive Coding (LPC) is one of the most important features based on auto-regression. LPC removes redundancy from a signal and attempts to determine the following values by linearly combining the previously known coefficients. LPC based features are used mainly for audio retrieval and segmentation [7].
2) *Short-Time Fourier Transform (STFT):* it is a time-frequency domain of a signal where such signal having a frequency on one axis and time on another. Generally, such a domain can be called as Time-Frequency Representation (TFR). The time-domain shows over the period, the variations in signal amplitude. On the other hand, the frequency domain gives frequency information about signal magnitude with no time information. To bridge this gap, TFR provides time and frequency information. The most efficient way to get a TFR is by using STFT [7].
3) *Chroma Related Features:* it can be computed from the logarithmic STFT of the sound signal. It is a powerful and interesting representation of music audio in which the entire spectrum is mapped into 12 bins that represent the 12 semitones (or Chroma) of the musical octave. It is also called chromatogram [7]. It is used in the accent recognition system as in [20].

4) *Tonality based Features:* In music, tonality organizes the notes of the musical scale. Tonality based features are also called prosodic features. The main tonality based features are Fundamental Frequency (F0), Harmonicity, and jitter [7], [21]. In general, F0 is a periodic waveform lowest frequency. It is also the first peek of the temporal autocorrelation function of the local normalized spectra. F0 is also known as voice pitch when a pitch is a sound wave perceptual property detectable by the human ear. The pitch may be measured as a frequency but it is not a physical property that is completely objective. It is a psychoacoustic, subjective attribute of sound. In the real world, people can identify the pitch of multiple sounds in real-time and separate each sound from a mixture of these. F0 or voice pitch has a significant parameter for classifying speech between male and female; for an adult male's voiced speech, F0 can vary between 85 and 180 Hz, and for an adult female, it may vary between 165 and 255 Hz. F0 is reported as above 200 Hz for kids. Hence, F0 is one of the most important features that are used in speaker gender and age estimation [12], [16]. Harmonicity features used to differentiate between the sounds of noise and the sounds of tones. To find sound periodicity, this feature uses autocorrelation function in the time or frequency domain. The ratio of the harmonic part of the signal to the rest of the signal can be calculated as Harmonic-to-Noise Ratio (HNR). It is widely used in sick voice analysis [22]. Jitter can be computed as F0 variations, that is, the mean absolute difference between consecutive speaking periods. It is used in determining vocal and non-vocal sounds [23], speaker recognition [13], speaker's gender and age estimation [14], [24].

5) *Spectrum Shape-based Features:* It is considered as one of the most important features in the audio signal. There are many types of spectrum-based features including spectral centroid, spectral roll-off, spectral spread, spectral skewness, spectral kurtosis, spectral slope, spectral decrease, spectral bandwidth, spectral flatness, spectral entropy, and spectral flux [7], [25]. Such features used in many areas like music mood classification [26], speech classification [27], Parkinson's disease detection from speech [28], speaker gender detection [21], [29], speaker age estimation [21], [30], and speaker emotion estimation [17].

The wavelet transform is another way to transform the time domain audio signal into a time-frequency representation. It computes the inner product of the signal with a member from family of wavelets. There are two types of wavelets: Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). The DWT has the capacity to extract information from non-stationary signals like audio. It overcomes the shortcomings of the STFT that provides uniform time-frequency resolution. DWT gives high time resolution and low frequency resolution for higher frequencies and high frequency resolution and low time resolution for lower frequencies. The approximations and detailed coefficients are generated by the wavelet transform that gives the information about a signal. These approximations and detailed coefficients are called as wavelet features [7].

By implementing the inverse Fourier transform for the signal spectrum logarithm, one can obtain the cepstral domain. There is power, real, complex, and phase cepstral. Of all of these, the power cepstrum is the most pertinent for the processing of speech signals. The cepstrum features are used primarily in speech recognition, pitch detection, speaker recognition, and speech enhancement [7]. The cepstrum features are shown in Fig. 4 and explained below.

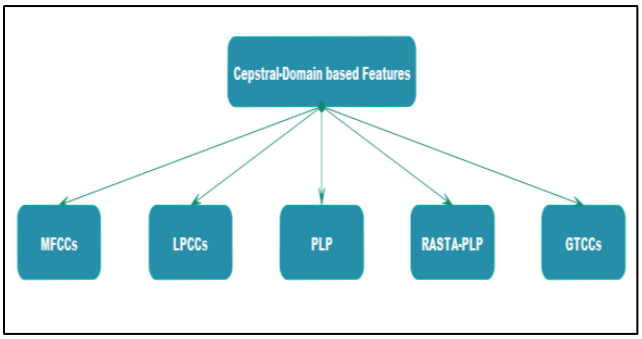

Fig. 4: Cepstral -Domain based Audio Features.

1) *Mel Frequency Cepstral Coefficients (MFCCs):* MFCC is originated from audio cepstral representation, it is recognized as one of the basic feature vectors in most acoustic pattern recognition problems. MFCC can symbolize the amplitude of speech signals concisely. In general, the resonance characteristics of the speaker's voice are affected by the shape of the vocal tract articulators, such as the teeth, nasal cavity, and tongue. This shape can give a precise illustration of the phoneme being formed if it is controlled precisely. The envelope of the short-time power spectrum can contain the vocal tract shape, hence, the MFCC purpose is to represent this envelope accurately. In MFCC, to closely mimic the human auditory system, the frequency bands should be equally spaced on the mel-scale. This makes MFCC as an audio signal processing key feature. In the existence of background noise, MFCC features are not exactly accurate and might not be well suited for generalization, and also it is sensitive to the environments [19], [21]. MFCCs has been widely used in different audio applications, such as speech recognition [31], speech enhancement [32], speaker recognition [33], [34], speaker authentication [35], voice activity detection [36], speaker gender detection [37], speaker age and gender estimation [38], [39], and speaker emotion recognition [40].

2) *Linear Prediction Cepstral Coefficients (LPCCs):* LPCC came from the transformation of previously mentioned LPC features into the cepstral domain; this may improve LPC efficiency because LPC is too sensitive to numerical precision. Transformation is done by taking DFT of the LPC logarithmic magnitude spectrum. LPCC is a typical example of representing features coefficients with less correlation, and that would be more efficient. Compared to LPC features,

LPCC features yield a lower error rate; however, LPCC has low vulnerability to noise [19]. It is used in the speech recognition system as in [41].

3) *Perceptual Linear Prediction (PLP) Cepstral Coefficients:* The PLP coefficients are based on three concepts, which are: critical band spectral resolution, intensity loudness power law, and equal-loudness curve. By using auto-regressive modeling preceded by perceptual processing, The PLP coefficients are produced from the LPC [19]. This feature used in several applications like speech recognition [42] and speaker emotion recognition [43].
4) *Relative Spectral Transform PLP (RASTA-PLP):* To incorporate the noise cancellation feature of the human auditory system, the RASTA-PLP features are presented. The RASTA-PLP features make PLP features robust to noise [21]. This hybrid feature is widely employed in gender classification [44], speaker age and gender estimation [21], and speech signals analysis [45].
5) *Gamma-Tone Cepstral Coefficients (GTCCs):* Noise reduction is one of the main problems of the speech recognition system. In recent years, GTCCs has shown noise robustness in numerous such system. It is based on gamma-tone filter banks; these filter banks provide a frequency-time representation of a signal name as cochleagram. Except that the 'Mel' filter bank is replaced by 'gamma-tone' filter bank, the extraction process of GTCCs is similar to that of MFCCs [7]. These groups of cepstral features are mainly used in speech recognition [46]. But it is also used in speech emotion recognition [47].

From low-level information, deep learning is considered as one of the most powerful and efficient techniques for extracting high-level features. These features' names as deep features can be extracted from various deep learning layers. The deep features could be extracted from any deep leaning model, like Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), unidirectional Long Short Term Memory network (LSTM), Bi-directional Long Short Term Memory (BLSTM) and other similar models [7].

In DNN, the deep features, extracted from lower layers, can be considered as speaker adapted features. On the other hand, deep features, extracted from DNN upper layers, can be considered as class-based discriminatory features. The deep features could be extracted from a DNN's bottleneck layers too. In CNN, there are three types of layers: convolution layers, pooling layers, and fully connected layers. By taking the spectrogram image of an audio signal, convolution layers apply some convolutional filters to produce feature map layers. For dimensionality reduction, pooling layers are conducted on feature map layers. Finally, global features are extracted from local map features by using fully connected layers. The deep features could be extracted from any of these three layers. RNNs are mainly used for sequential knowledge modeling, it can memorize information internally. There are two types of memory elements in RNNs: Unidirectional or bi-directional LSTM. The information processing flow is forward in unidirectional LSTM, while it is both forward and backward in BLSTM. The deep features could be extracted from any LSTM-RNN deep layers [7].

In audio signal processing, deep features have been used in speech recognition [48], gender recognition [49], speaker age and gender estimation [50], classification of acoustic scene [51], emotion recognition [52], spoofing detection [53], audio-video analysis [54], and speech-music discrimination [55].

B. Dimensionality Reduction

Predominantly, there is some redundancy in the extracted high-dimensional features. Subspace learning can be used to eliminate these redundancies by further processing extracted features to reflect their semantic information better. Subspace learning aims at revealing and preserving the original signals' essential features by mapping high-dimensional features space into a low-dimensional feature space. Representative methods include Principle Component Analysis (PCA), Locality Preserving Projections (LPP), and Linear Discriminant Analysis (LDA) [1]. If one has n-dimensions features vector to be reduced, PCA tries to search for k-dimensions that can be efficiently used for representing these features, where ($k \leq n$). This results in dimensionality reduction by projecting the high dimensional features onto a much smaller space. By creating an alternative smaller set of variables, PCA combines the essence of feature dimensions. PCA often reveals relationships that were not suspected before and thus allows interpretations that would not normally result [56].

The aim of LPP, which is one of the manifold learning methods is to preserve the original data intrinsic geometry structure in addition to keeping locality relationships after projection by making the samples present in a neighborhood. To characterize the training set neighborhood relationship, LPP constructs one affinity graph. After that, it seeks a low-dimensional embedding to preserve the local structure and intrinsic geometry [1].

LDA uses cases where the within-class frequencies are unequal, and their performance has been investigated on data collected randomly. LDA ensures maximum separability in any dataset by maximizing the ratio of the between-class variance to the within-class variance. LDA does data classification, while PCA does more of feature classification. This can be considered as the difference between them. In PCA, when transformed into a different space, the location and shape of the original data sets change, whereas LDA does not change the location but only attempts to provide more class separability and draw a region of decision between the given classes [57].

C. *Semantics Understanding*

For social robots, semantic understanding tasks must be conducted after the features extraction process. In general, various tasks involve different methods of semantic understanding for voice-based perception; for example, speech recognition, speaker recognition, speaker gender detection, speaker age estimation, and speaker localization. These tasks have several HRI applications.

1) *Speech Recognition:* This task is considered as the ability to recognize spoken language words or phrases

and then convert them to a machine-readable format. Hence, the process of recognizing the person's spoken words are automatically based on speech signal information called Automatic Speech Recognition (ASR) [58], [59]. The development of ASR systems began with simple digit recognition systems, then isolated words, continuously spoken words in a silent environment, up to the recognition of spontaneous speech in a noisy environment. There were three important moments in the development of ASR systems: introduction of MFCC, the introduction of statistical methods such as Hidden Markov Model (HMM), and Gaussian Mixture Model (GMM), and introduction of DNN [60].

2) *Speaker Recognition:* Speaker recognition is a biometric authentication process, in which human voice characteristics are used as the attribute. The typical speaker recognition system consists of a modeling scheme to characterize the speaker's voice features using a statistical approach and a classification scheme for the unknown utterance characteristics. Generally, such a recognition system can be classified into two processes, which are speaker verification and speaker identification. The main difference between these two categories is that the speaker verification performs a binary decision to verify the speaker's identity whereas speaker identification performs multiple decisions and it consists of comparing the voice of the person speaking to some reference templates in an attempt to identify the speaker [34], [61].

3) *Speaker Gender Detection*: Speech signals can contain some information in addition to speaker identity, like speaker age, speaker gender, and speaker emotional state. Speaker gender is one of the most specific speech information; if one can identify the gender of the speaker, this can be useful in many fields. Gender dependent speech recognition system shows a decrease in word error rate compared with the basic system. On the other hand, the prediction of other speaker traits, such as age and emotion, can be improved if gender identification was conducted first. In general, gender identification is important for ever more natural and personalized dialog systems [29], [37].

4) *Speaker Age Estimation:* Speech signal paralinguistic information, like speaker identity, speaker emotional state, speaker gender, and speaker age, can guide HRI systems to automatically adapt to different user needs. One of the most challenging tasks is to identify the age of speakers from a short speech utterance. Age information helps to adapt interactive voice response systems to the user and that can give more natural HRI. The speed of the speech synthesis system also may change depending on the user age. A better understanding of speaker age will help improve voice analysis for a variety of applications including computers, cell phones, and robots. While human listeners are generally believed to be able to judge a speaker's age within ±10 years, few proposals can accomplish this task yet. The use of gender detection before the process of age estimation has a major impact on improving results, hence most of the proposals to estimate the speaker age are preceded by gender detection [24], [61].

5) *Speaker Localization:* In HRI, this method is usually used to locate a speaker so that the robot's attention is directed to this speaker. Initially, the speaker's audio signal is received through the robots' microphones, after that the temporal shifts between audio signals are calculated by the cross-power spectrum phase. Finally, the speaker's location is determined by the corresponding microphone by calculation the time delay of arrival from the temporal shift [1].

D. Text-To-Speech Synthesis

It is a system that converts normal language text into speech. There are a lot of differences among human and machine speech production, but the increase in the ability of machine learning paradigms to simulate human speech production mechanism will increase the accuracy and naturalness of TTS. The first full TTS system for English was introduced in 1968. It was an articulatory-based system that could perform text analysis and determine pauses in the text using a sophisticated parser. However, it was not until concatenative synthesizers were invented, that TTS gained widespread usage. The idea of concatenative TTS is to concatenate appropriate parts of a prerecorded database. This method is extremely inflexible in terms of changing the speaking style or the voice of the speaker; it requires a whole new database to be recorded and annotated. DNN was effectively used for mapping input linguistic features to output acoustic features, enabling nonlinear mappings. Although DNN with several hidden layers and sigmoid or tangent hyperbolic activations are sufficient for the production of intelligible and natural-sounding synthetic speech, the introduction of LSTM units has brought further improvement into the quality of synthesized speech. DNNs have not only just enabled generating synthetic speech of high quality but also introduced many possibilities for the production of speech in different voices and speaking styles [60].

IV. REPRESENTATIVE VOICE-BASED SOCIAL ROBOTS METHODS AND TECHNIQUES

Despite all the challenges encountered when implementing social robots in real-world applications, there are some social robots developed to help our daily lives. Some representative voice-based social robot methods and techniques are presented in this section. Table 1. summarizes these works from three aspects, namely feature type, semantic understanding type, and core algorithm.

Breazeal et al. [62] proposed a system to identify the intention of emotional speech directed to the robot, where social learning is built between the robot and its human caregiver. That leads to the human caregiver's ability to directly modulate the emotional state of the robot through verbal communication. They used GMM as a classifier, and pitch and energy as a speech feature.

TABLE 1.
SOCIAL ROBOTS VOICE-BASED PERCEPTION METHODS AND TECHNIQUES.

Authors	*Year*	*Type of Feature*	*Semantic Understanding Task*	*Core algorithm*
Betkowska et al. [65]	2007	Cepstral Based	Speech Recognition	FHMM
Granata et al. [68]	2010	n/a	Speech Recognition	n/a
Strait et al. [69]	2015	n/a	Speech Recognition	n/a
Poncela et al. [72]	2014	Julius	Speech Recognition	Julius
Lee et al. [38]	2012	MFCC, LPCC	Gender and Age Group Recognition	SVM, C4.5 Decision Tree
Kim et al. [14]	2007	MFCC	Gender and Age Recognition	GMM
Schmitz et al. [67]	2009	Time Delay	Sound Localization	Beam-Forming
Niculescu et al. [71]	2011	Pitch	Voice Attraction	n/a
Breazeal et al. [62]	2002	Pitch, Energy	Emotion Recognition	GMM
Ruvolo et al. [63]	2008	STBF-based	Emotion Recognition	Boosting
Austermann et al. [66]	2005	Prosody Based	Emotion Recognition	Fuzzy Logic
Tahon et al. [70]	2016	Multi-level	Emotion Recognition	SVM
Kraft et al. [64]	2005	MFCC	Sound Classification	HMM, GMM

Ruvolo et al. [63] presented an auditory mood detection approach, which was raised from their experience immersing social robots in classrooms. Their approach to recognize auditory categories has been inspired by the machine learning methods used in computer vision. After converting the auditory signal into a sonogram, which is an image of the acoustic signal, they use a new collection of enormous lightweight spatio-temporal filters. Their approach showed good performance in a human speech regarding the problem of emotional state recognition.

Kraft et al. [64] presented a robot used in a kitchen environment for a sound classification task. They used a temporal extension of Independent Component Analysis (ICA) as a feature extraction method, while, they used GMM and HMM as classifiers in their approach.

Betkowska et al. [65] highlighted the speech recognition problem in the existence of nonstationary sudden noise, which will most probably happen in home environments. They used the Factorial Hidden Markov Model (FHMM) architecture which was developed from a sudden-noise and a clean-speech HMM. To assess their proposed method, they used a personal robot to record their database in home environments. Their experiments have shown that their proposed method is effective under noisy conditions.

Austermann et al. [66] presented a speech-based emotion recognition approach by using a robot. They used fuzzy logic for a prosody analysis in natural speech. In their approach, they realized a speaker-dependent mode as well as a speaker-independent mode because the robot can communicate with well-known humans or with unknown humans. Based on a training database, their approach selects the most significant features automatically from a set of twenty analyzed features. According to their results, this is important, since the set of significant features differs significantly between the distinguished emotions.

Schmitz et al. [67] stated that one of the basic components for natural HRI is an audio system. Hence, they presented an audio perception module with a 6 microphone array for sound localization tasks. This task is implemented using the beam-forming algorithm which assigns the probability of sound to each virtual measuring point in the robot environment. Ultimately, the filtering module uses the beam-former output to produce sector maps, which minimize

a large amount of data to the localization process essential information.

Lee et al. [38] presented a comparative study in gender and age group recognition for HRI applications. They concentrate on audio-based techniques from multichannel microphones and soundboards equipped with robots, among different HRI components. For comparative purposes, they carry out the performance comparison of LPCC and MFCC in the feature extraction step, Quinlan's earlier Decision Tree (C4.5) method, and Support Vector Machine (SVM) in the classification step. Finally, they deal with the usefulness of HRI recognition for gender and age groups in-home service robot environments.

Kim et al. [14] presented speech-based gender and age recognition for HRI. They use MFCC as a feature and GMM as a classifier to assign service applications for robots that can satisfy users by providing services tailored to the specific needs of specific groups of users, including females and males, adults, and children. The main aim of their work is to extract voice-based gender and age information to classify these characteristics for HRI.

Granata et al. [68] presented a multimodal approach for HRI dedicated to the elderly. To facilitate the robot's use, they used clear and simple voice and graphical-based interfaces. Despite voice interactions, seen as the most common way in which people express their needs, they claimed older users were less likely to communicate by speech. Hence, they propose a multimodal approach to make older users interact comfortably with the machine. Their voice-based interface consists of an automatic speech recognition interpreting the human voice, a dialog manager selecting the appropriate response, and a TTS module synthesizing the chosen response.

Strait et al. [69] stated that a robot using politeness modifiers in its speech perceives human interaction more favorably than a robot using direct speech. They have found that polite speech by a robot is influenced by other factors as well, such as gender and age; for example, females tend to interact with robots through polite speech more than men. Besides, the elderly and children tend to have this type of interaction as well.

Tahon et al. [70] presented voice-based detection of emotions in the context of HRI for both the elderly and children. Their system is based on multi-level processing of the audio cues. They estimate their performance of the emotion detection system in cross-corpus conditions since their models are built with data recording, and the system will test real-life data.

Niculescu et al. [71] presented a social robot receptionist evaluation in terms of the robot voice pitch effect. In their experiments, they used two types of robots, one with low pitch; the other with a high pitch. Their results show that robots with high pitch perceived more attraction in terms of behavior, personality, and voice. Their study would like to highlight the importance of voice in general and voice pitch in social robot design in particular.

Poncela et al. [72] presented an HRI for teleoperation a robotic platform using user speech. Hence, they designed a user-dependent acoustic model based on Spanish speakers to teleoperated a robot with a collection of verbal commands. Their results have been successful, both in robot navigation under the user's voice commands, and a high recognition rate.

V. DISCUSSIONS

As described in this review, the voice-based perception for HRI can be primarily included in three steps: feature extraction, dimensionality reduction, and semantic understanding. On the other hand, the social robot's action is represented by a speech response through the use of the TTS method.

In the feature extraction step, the voice-based feature can be extracted from a variety of domains such as time, frequency, cepstral, or deep domain. For each voice-based semantic understanding tasks, there is a certain type of feature that suits it. For instance, recently, speaker gender detection and age estimation are among the most common areas of research. According to the literature review, the most discriminatory and suitable features have been proven for machine learning techniques in such areas that include ZCR, Shimmer, STE, F0, Jitter, Spectral Features, MFCC, and Deep Features. Hence, the social robots' semantic understanding task is an important factor in selecting and developing a voice-based features extraction method.

In the dimensionality reduction step, subspace learning can be used to eliminate the redundancies of high-dimensional features by further processing extracted features to reflect their semantic information better. PCA has been considered as the most common method for unsupervised reduction, while LDA is the most common method for supervised reduction.

In the semantic understanding step, some voice-based semantic understanding tasks have been explained. For each application, there is a certain type of voice-based semantic understanding task that suits it, especially for social robots. The voice-based semantic understanding mainly relies either on classification methods or regression methods. For instance, between all of the voice-based semantic understanding tasks, one of the most popular research areas recently is speaker gender detection and age estimation, which is used classification method for gender detection, and, regression method for age estimation. According to the literature review, the most discriminatory and suitable machine learning methods have been proven that for this semantic understanding task include SVM, GMM, HMM, and Boosting. It can be seen that how to develop voice-based semantic understanding tasks mainly depends on the application area. Hence, the researchers must choose carefully the machine learning method for the semantic understanding task that suits their application.

On the other hand, DNN-based synthetic speech has already attained such a quality that is difficult to distinguish from human speech. With the flexibility of changing speaker and style.

Despite all the efforts made and all research developed in the voice-based HRI field, some challenges remain to be addressed in future works:

1) Numerous features can be extracted from an audio signal, each of which suits one or more specific applications. However, choosing the appropriate feature

that gives the best performance for a specific application remains a challenge.

2) One way to improve semantic understanding task performance is feature fusion. However, how to integrate them efficiently remains a challenge.

3) There are great differences between the laboratory environment in which the system was built, and the actual environment in which the system was used. How to implement a robust system in real environments still a challenge.

4) There are numerous semantic understanding methods based on voice, each of which fits a specific application. How to choose the appropriate semantic understanding method for a specific application while ensuring that the appropriate features are used, remains something that needs to be investigated further.

5) Supervised machine learning algorithms can be replaced by Reinforcement-based machine learning algorithms, which will, in turn, bring about progress in areas where large data sets are not available, as is the case in speech synthesis.

6) Lots of researchers working on social robots' perception tasks have accomplished promising performance. However, how to improve these results remains to be investigated more to obtain high autonomy and lower computational costs.

VI. Conclusions

Among all kinds of interactions between robots and humans, voice-based interaction is the most realistic. This work provided a comprehensive review of the voice-based interface for HRI systems from three aspects: feature extraction, dimensionality reduction, and semantic understanding. Besides, representative voice-based social robot methods and techniques have been reviewed. The main challenges that face researchers when building such systems have been given. Moreover, recommendations and trends for future works and further explorations of this area have been also outlined.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] H. Yan, M. H. Ang, and A. N. Poo, "A survey on perception methods for human–robot interaction in social robots," *Int. J. Soc. Robot.*, vol. 6, no. 1, pp. 85–119, 2014.

[2] P. Tsarouchi, S. Makris, and G. Chryssolouris, "Human–robot interaction review and challenges on task planning and programming," *Int. J. Comput. Integr. Manuf.*, vol. 29, no. 8, pp. 916–931, 2016.

[3] N. Lubold, E. Walker, and H. Pon-Barry, "Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2016, pp. 255–262.

[4] M. A. Goodrich and A. C. Schultz, *Human-robot interaction: a survey*. Now Publishers Inc, 2008.

[5] H. A. Yanco and J. Drury, "Classifying human-robot interaction: an updated taxonomy," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, 2004, vol. 3, pp. 2841–2846.

[6] M. Ahmad, O. Mubin, and J. Orlando, "A systematic review of adaptivity in human-robot interaction," *Multimodal Technol. Interact.*, vol. 1, no. 3, p. 14, 2017.

[7] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Appl. Acoust.*, vol. 158, p. 107020, 2020.

[8] T. H. Zaw and N. War, "The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 2017, pp. 1–5.

[9] X. Yang, B. Tan, J. Ding, J. Zhang, and J. Gong, "Comparative study on voice activity detection algorithm," in *2010 International Conference on Electrical and Control Engineering*, 2010, pp. 599–602.

[10] Y. Korkmaz, A. Boyacı, and T. Tuncer, "Turkish vowel classification based on acoustical and decompositional features optimized by Genetic Algorithm," *Appl. Acoust.*, vol. 154, pp. 28–35, 2019.

[11] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and a da b oost for music classification," *Mach. Learn.*, vol. 65, no. 2–3, pp. 473–484, 2006.

[12] O. T.-C. Chen and J. J. Gu, "Improved gender/age recognition system using arousal-selection and feature-selection schemes," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 148–152.

[13] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth annual conference of the international speech communication association*, 2007.

[14] 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, pp. 122–126.

[15] F. Metze *et al.*, "Comparison of four approaches to age and gender recognition for telephone applications," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 2007, vol. 4, pp. IV–1089.

[16] M. Li, C.-S. Jung, and K. J. Han, "Combining five acoustic level modeling methods for automatic speaker age and gender recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," in *2006 14th European Signal Processing Conference*, 2006, pp. 1–5.

[18] M. P. Michalowski, S. Sabanovic, and H. Kozima, "A dancing robot for rhythmic social interaction," in

Proceedings of the ACM/IEEE international conference on Human-robot interaction, 2007, pp. 89–96.
[19] S. A. Alim and N. K. A. Rashid, "Some commonly used speech feature extraction algorithms," *From Nat. to Artif. Intell. Appl.*, 2018.
[20] K. Mannepalli, P. N. Sastry, and M. Suman, "Accent Recognition System Using Deep Belief Networks for Telugu Speech Signals," in *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, 2017, pp. 99–105.
[21] B. D. Barkana and J. Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," *Appl. Acoust.*, vol. 98, pp. 52–61, 2015.
[22] J.-W. Lee, S. Kim, and H.-G. Kang, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5969–5973.
[23] Y. V. S. Murthy and S. G. Koolagudi, "Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations," in *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2015, pp. 1271–1276.
[24] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel, "Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 1975–1985, 2011.
[25] A. G. Jondya and B. H. Iswanto, "Indonesian's traditional music clustering based on audio features," *Procedia Comput. Sci.*, vol. 116, pp. 174–181, 2017.
[26] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 14, no. 1, pp. 5–18, 2005.
[27] A. I. Al-Shoshan, "Speech and music classification and separation: a review," *J. King Saud Univ. Sci.*, vol. 19, no. 1, pp. 95–132, 2006.
[28] T. Tuncer and S. Dogan, "A novel octopus based Parkinson's disease and gender recognition method using vowels," *Appl. Acoust.*, vol. 155, pp. 75–83, 2019.
[29] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proceeding of speech prosody*, 2016, pp. 84–88.
[30] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and gender recognition based on multiple systems-early vs. late fusion," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
[31] E. S. Wahyuni, "Arabic speech recognition using MFCC feature extraction and ANN classification," in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2017, pp. 22–25.
[32] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 18, no. 7, pp. 1692–1707, 2010.
[33] A. K. Abdul-Hassan and I. H. Hadi, "Speaker Classification using DTW and A Proposed Fuzzy Classifier," *J. Al-Qadisiyah Comput. Sci. Math.*, vol. 11, no. 4, p. Page-17, 2019.
[34] C. Sunitha and E. Chandra, "Speaker Recognition using MFCC and Improved Weighted Vector Quantization Algorithm," *Int. J. Eng. Technol.*, vol. 7, pp. 1685–1692, Nov. 2015.
[35] A. K. Abdul-Hassan and I. H. Hadi, "A Proposed Authentication Approach Based on Voice and Fuzzy Logic," in *Recent Trends in Intelligent Computing, Communication and Devices*, Springer, 2020, pp. 489–502.
[36] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using MFCC features and support vector machine," in *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, 2007, vol. 2, pp. 556–561.
[37] G. ALIPOOR and E. SAMADI, "Robust Speaker Gender Identification Using Empirical Mode Decomposition-Based Cepstral Features," *vol*, vol. 7, pp. 71–81.
[38] M.-W. Lee and K.-C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 12, 2012.
[39] S. Safavi, M. Russell, and P. Jančovič, "Automatic speaker, age-group and gender identification from children's speech," *Comput. Speech Lang.*, vol. 50, pp. 141–156, 2018.
[40] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 474–477.
[41] H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in Speech Recognition System," in *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, 2016, pp. 498–502.
[42] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *Int. J. Adv. Res. Eng. Technol.*, vol. 1, no. 6, pp. 1–4, 2013.
[43] M. Glodek *et al.*, "Multiple classifier systems for the classification of audio-visual emotional states," in *International Conference on Affective Computing and Intelligent Interaction*, 2011, pp. 359–368.
[44] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech," in *2006 International Conference on Machine Learning and Cybernetics*, 2006, pp. 3376–3379.
[45] M. A. A. Zulkifly and N. Yahya, "Relative spectral-perceptual linear prediction (RASTA-PLP) speech signals analysis using singular value decomposition (SVD)," in *2017 IEEE 3rd International Symposium in Robotics and Manufacturing Automation (ROMA)*, 2017, pp. 1–5.
[46] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency," *Speech Commun.*, vol. 53, no. 5, pp. 707–715, 2011.
[47] G. K. Liu, "Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech," *arXiv Prepr. arXiv1806.09010*, 2018.
[48] M. H. Rahmani, F. Almasganj, and S. A. Seyyedsalehi,

"Audio-visual feature fusion via deep neural networks for automatic speech recognition," *Digit. Signal Process.*, vol. 82, pp. 54–63, 2018.
[49] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Appl. Acoust.*, vol. 156, pp. 351–358, 2019.
[50] A. A. Mallouh, Z. Qawaqneh, and B. 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, pp. 2581–2593, 2017.
[51] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu, and X. Feng, "Acoustic scene classification using deep audio feature and BLSTM network," in *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, 2018, pp. 371–374.
[52] A. M. Badshah *et al.*, "Deep features-based speech emotion recognition for smart affective services," *Multimed. Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, 2019.
[53] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Commun.*, vol. 85, pp. 43–52, 2016.
[54] N. Takahashi, M. Gygli, and L. Van Gool, "Aenet: Learning deep audio features for video analysis," *IEEE Trans. Multimed.*, vol. 20, no. 3, pp. 513–524, 2017.
[55] M. Papakostas and T. Giannakopoulos, "Speech-music discrimination using deep visual feature extractors," *Expert Syst. Appl.*, vol. 114, pp. 334–344, 2018.
[56] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
[57] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," in *Institute for Signal and information Processing*, 1998, vol. 18, no. 1998, pp. 1–8.
[58] O. Mubin, J. Henderson, and C. Bartneck, "You just do not understand me! Speech Recognition in Human Robot Interaction," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 637–642.
[59] S. D. Dhingra, G. Nijhawan, and P. Pandit, "Isolated speech recognition using MFCC and DTW," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, vol. 2, no. 8, pp. 4085–4092, 2013.
[60] V. Delić *et al.*, "Speech technology progress based on new machine learning paradigm," *Comput. Intell. Neurosci.*, vol. 2019, 2019.
[61] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Commun.*, vol. 54, no. 4, pp. 543–565, 2012.
[62] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Auton. Robots*, vol. 12, no. 1, pp. 83–104, 2002.
[63] P. Ruvolo, I. Fasel, and J. Movellan, "Auditory mood detection for social and educational robots," in *2008 IEEE International Conference on Robotics and Automation*, 2008, pp. 3551–3556.
[64] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, "Temporal ICA for classification of acoustic events ia kitchen environment," in *Ninth European Conference on Speech Communication and Technology*, 2005.
[65] A. Betkowska, K. Shinoda, and S. Furui, "Robust speech recognition using factorial HMMs for home environments," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 20593, 2007.
[66] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Prosody based emotion recognition for MEXI," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 1138–1144.
[67] N. Schmitz, C. Spranger, and K. Berns, "3D audio perception system for humanoid robots," in *2009 Second International Conferences on Advances in Computer-Human Interactions*, 2009, pp. 181–186.
[68] C. Granata, M. Chetouani, A. Tapus, P. Bidaud, and V. Dupourqué, "Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders," in *19th International Symposium in Robot and Human Interactive Communication*, 2010, pp. 785–790.
[69] M. Strait, P. Briggs, and M. Scheutz, "Gender, more so than age, modulates positive perceptions of language-based human-robot interactions," in *4th international symposium on new frontiers in human robot interaction*, 2015, pp. 21–22.
[70] M. Tahon, A. Delaborde, and L. Devillers, "Real-life emotion detection from speech in human-robot interaction: Experiments across diverse corpora with child and adult voices," 2011.
[71] A. Niculescu, B. Van Dijk, A. Nijholt, and S. L. See, "The influence of voice pitch on the evaluation of a social robot receptionist," in *2011 International Conference on User Science and Engineering (i-USEr)*, 2011, pp. 18–23.
[72] A. Poncela and L. Gallardo-Estrella, "Command-based voice teleoperation of a mobile robot via a human-robot interface," *Robotica*, vol. 33, no. 1, p. 1, 2015.