International Journal of Social Robotics (2024) 16:659–670
https://doi.org/10.1007/s12369-022-00913-x

A Scoping Review of the Literature On Prosodic Elements Related to Emotional Speech in Human-Robot Interaction

Norina Gasteiger[1,2] · JongYoon Lim[1] · Mehdi Hellou[1,3] · Bruce A. MacDonald[1] · Ho Seok Ahn[1]

Accepted: 3 August 2022 / Published online: 20 October 2022

Abstract

Background Sentiment expression and detection are crucial for effective and empathetic human-robot interaction. Previous work in this field often focuses on non-verbal emotion expression, such as facial expressions and gestures. Less is known about which specific prosodic speech elements are required in human-robot interaction. Our research question was: what prosodic elements are related to emotional speech in human-computer/robot interaction?

Methods The scoping review was conducted in alignment with the Arksey and O'Malley methods. Literature was identified from the SCOPUS, IEEE Xplore, ACM Digital Library and PsycINFO databases in May 2021. After screening and de-duplication, data were extracted into an Excel coding sheet and summarised.

Results Thirteen papers, published from 2012 to 2020 were included in the review. The most commonly used prosodic elements were tone/pitch (n = 8), loudness/volume (n = 6) speech speed (n = 4) and pauses (n = 3). Non-linguistic vocalisations (n = 1) were less frequently used. The prosodic elements were generally effective in helping to convey or detect emotion, but were less effective for negative sentiment (e.g., anger, fear, frustration, sadness and disgust).

Discussion Future research should explore the effectiveness of commonly used prosodic elements (tone, loudness, speed and pauses) in emotional speech, using larger sample sizes and real-life interaction scenarios. The success of prosody in conveying negative sentiment to humans may be improved with additional non-verbal cues (e.g., coloured light or motion). More research is needed to determine how these may be combined with prosody and which combination is most effective in human-robot affective interaction.

Keywords affective computing · speech · HRI · robotics · social robots · sentiment

1 Introduction

The concept of 'affective computing' suggests that computers may be capable of detecting emotion, responding appropriately and expressing emotion [1]. In this sense, computer systems incorporate human-like emotional intelligence and empathy [2]. In the field of affective computing, emotion is defined as a systems' subjective interpretation of meaningful events [3]. Within robotics, sentiment analysis and emotion understanding are essential to developing longer-term relationships and rapport with human users, especially to maintain interest when the novelty of engagement wears off [4].

Sentiment expression is widely understood to be multimodal, requiring both non-verbal and verbal efforts that are mutually understood by and align with the interactive parties. Non-verbal parameters include facial expressions and head movements (i.e., tilting). Verbal expression of emotion is more complicated and includes semantics (content of speech) and prosodic cues which impact meaning (i.e., intonation, pitch, volume/energy, pauses or speed) [5].

Through changes in verbal and non-verbal parameters, various emotions can be expressed (and ultimately detected). Previous work on cross-cultural human-human interaction has reported that many of these speech features can be used

✉ Ho Seok Ahn
hs.ahn@auckland.ac.nz

1 Department of Electrical, Computer and Software Engineering, University of Auckland, 1142 Auckland, New Zealand

2 School of Health Sciences, The University of Manchester, Manchester, UK

3 Département d'Informatique, Facultés des Sciences et Ingénieries, Sorbonne Université, Paris, France

universally, to correctly identify emotion [6]. Specifically, humans identify joy/happiness by a rapid speaking rate, higher pitch and larger pitch range, while sadness can be detected through a slower speaking rate, lower average pitch and more narrow range [7]. Anger and fear both have a faster speaking rate, but anger has a rising pitch contour, while fear may have more varied loudness [7].

The 'Big Six' or 'Big Eight' categories are often used to understand and distinguish between main emotions. The Big Six include: anger, disgust, fear, joy/happiness, sadness, and surprise [8–10]. Plutchik identified eight emotions, adding anticipation and trust to the Big Six [11]. These were also understood to have opposing emotions (e.g., happiness/joy and sadness) and range in intensity. As a result, trends measuring emotion have resulted in Russell's Circumplex Model of Affect, whereby emotions revolve around arousal and valence [12]. Computational modelling of sentiment has aimed to detect these continuous values [13, 14]. Researchers have also used a form of multiclass classification, whereby the classes include the categories as identified as above, or simply: positive, negative and neutral [9, 10, 15].

However, sentiment expression and detection are extremely complex. This is because prosodic factors such as intonation do not consist of single independent systems, but are a product of the amalgamation of various features, including tone, loudness, tempo, rhythm and pitch range and contour [16]. Combining verbal expressions as such with non-verbal expressions in systems can be helpful, but also more complicated. This is because emotional incongruence may occur, whereby the different modalities indicate differing emotions [17]. For example, an individual may smile while presenting bad news, or demonstrate sarcasm with a serious facial expression. While humans can typically navigate this incongruence, on-going research is dedicated to attempting to understand how [18, 19]. Nuances such as these are important to understand when attempting to design computer and robotic systems that can detect and express sentiment.

Previous research on human-robot interaction has focussed on non-verbal expressions of sentiment, including facial and body expressions [20, 21]. Less has been conducted purely on prosodic factors. However, this has included augmenting 'robotic' voices with forms of prosody in an attempt to convey sentiment [5]. A gap in knowledge remains on which prosodic factors are required for successful sentiment expression and detection in human-robot interaction.

This review is part of an over-arching project that seeks to develop a sentiment analyser that can be implemented on a robot. Our previous work has included the development of a coverage-based sentiment and sub-sentence extraction system that estimates a span of input text and recursively feeds this information back to the networks for sentiment identification [4]. Twenty-four ablation studies were conducted and showed promising results. Our next step, and the aim of this review, is to understand how emotional speech is expressed or detected within existing robotic systems, with a focus on prosody.

2 Methods and Methodology

We conducted a scoping review, as this method seeks to explore and synthesise the available literature, as well as to map relevant ideas and concepts in regard to the research topic [22, 23]. As in other reviews, scoping reviews are conducted systematically and transparently, but like narrative and descriptive reviews, cover the breadth (not the depth) of research by summarizing previous knowledge [24, 25].

The methods and procedures of this scoping review aligned with the established five-step process proposed by Arksey and O'Malley [24]: (1) identifying the research question, (2) identifying relevant studies, (3) selecting studies, (4) charting the data, and (5) collating, summarizing, and reporting the results. We also report this review in alignment with the PRISMA-ScR guideline [26].

2.1 Identifying the Research Question

The purpose of this project is to develop a sentiment analyser that can be implemented on a robot. Therefore, we wanted to understand how emotional speech is expressed or identified within existing robot systems, such as by using tone, speed or pitch. This helped us to consequently develop the research question: What prosodic elements are related to emotional speech in human-computer/robot interaction?

2.2 Identifying Relevant Studies

Four engineering and social science databases were searched on the 5th May 2021. These included SCOPUS, IEEE Xplore, ACM Digital Library and PsycINFO.

Keywords (usually synonyms) were separated by the Boolean operators 'AND' and 'OR'. The search strategy included the following: (emotion OR sentiment) AND (speech OR verbal OR tone OR pitch) AND (expression OR identification) AND (experiment OR evaluation) AND robot.

To focus the review on more recent and current state-of-the-art methods, we decided to limit our search to the last 10 years, covering 2011 to 2021. Other limits included being published in English, focussing on humans (not animals) and the full-text being available. All items had to include an experiment/evaluation or research component, as well

Table 1 Examples of the search syntax used in two of the databases

Database	Search syntax
SCOPUS	(TITLE-ABS-KEY (emotion OR sentiment) AND TITLE-ABS-KEY (speech OR verbal OR tone OR pitch) AND TITLE-ABS-KEY (expression OR identification) AND TITLE-ABS-KEY (experiment OR evaluation) AND TITLE-ABS-KEY (robot)) AND PUBYEAR > 2010 AND PUBYEAR < 2022 AND (LIMIT-TO (LANGUAGE, "English"))
PsycINFO	1. ((emotion or sentiment) and (speech or verbal or tone or pitch) and (expression or identification) and (experiment or evaluation) and robot).af. 2. limit 1 to (full text and english language and yr="2011 -Current")

as focus on prosody, rather than semantics. Items on multimodal emotion (i.e., facial expression and speech) were only included if the content on speech could be separated and was discussed in sufficient detail. Table 1 exemplifies the search syntax used in two of the databases.

2.3 Selecting the Studies

We created an Excel sheet, to document the searches. This document included the dates of each search, the databases searched and the literature identified through the searches. A two-step process helped to screen the literature for eligibility. This included first reading the abstracts and titles of the items, and removing those that did not meet the eligibility criteria. Duplicate studies between the searches were also removed in this step. The second step consisted of downloading and reading the full-text items and identifying reasons for exclusion. Items that passed the full-text screening stage were included in the review. The screening process was presented in a PRISMA diagram [27].

2.3.1 Charting the Data and Collating, Summarizing, and Reporting the Results

A second Excel sheet formed our coding framework, into which we extracted relevant data from each study. This included the following: title of the publication, first authors surname and publication date (year), publication type (journal or conference paper), study setting and country, robot/system, purpose of the robot/system, speech detection/expression, prosodic factors and description of them, study participants, evaluation/study method and outcome.

Data from the coding sheet were summarised and presented in a manner which best answered the research question.

3 Results

The database search yielded 1,889 results. Twenty-seven duplicates were removed, and another 1,806 were excluded during the first screening process. During the second screening, 43 studies were excluded with reasons. These included not being on speech (n = 12), not focusing on prosody (n = 11), not reporting on sentiment (n = 7) or experimental results (n = 7), not including anything on human-robot interaction (n = 3) and being published in languages other than English (n = 3). Consequently, a total of 13 publications were included in the review. The PRISMA diagram [27] in Fig. 1 demonstrates the search and screening process.

3.1 Characteristics of the Included Literature

The literature was published from 2012 [28] to 2020 [29, 30]. The majority of the studies were presented at conferences and subsequently published as full-text conference papers [5, 30–37]. Only two were published as journal papers [28, 29]. One study appeared to be published as a journal article and as a conference paper [38, 39], hence is reported together.

The studies were conducted in France [32, 35], Japan [29] and Spain and Poland [5]. Nine did not state the specific country location [28, 30, 31, 33, 34, 36–39]. The settings of the research were most commonly places of education (schools and universities) [29, 32, 33, 35, 38, 39] or online [30, 31]. One was conducted in a Alzheimer's Center and hospital [5] and four did not state the specific setting [28, 34, 36, 37].

Overall, the studies were mostly identified to be experimental in nature. Two appeared to be observational, in which differences in prosodic elements between synthesised and natural human speech were explored [37] and responses to different pitches were observed [29]. A further two studies collected mixed methods data, such as qualitative perceptions (via a survey or interviews) in addition to quantitative measures [31, 32].

Across the studies, the sample sizes varied from three [37] to 300 participants [36], including children, adults and students. Of the studies (n = 10) that specified a sample size, a total of 774 participants were included (mean: 77.4). The studies are further summarised in Table 2.

3.2 The Systems

Most of the studies used robots, including NAO [30], ERICA [38, 39], Pepper [29, 31], the Survivor Buddy robot [33], RAMCIP [5], ALICE [35] and Hobbit [31]. Three studies used systems [28, 36, 37], such as the VOICEROID 2 Yuduki Yukari speech synthesiser system [37]. Only one

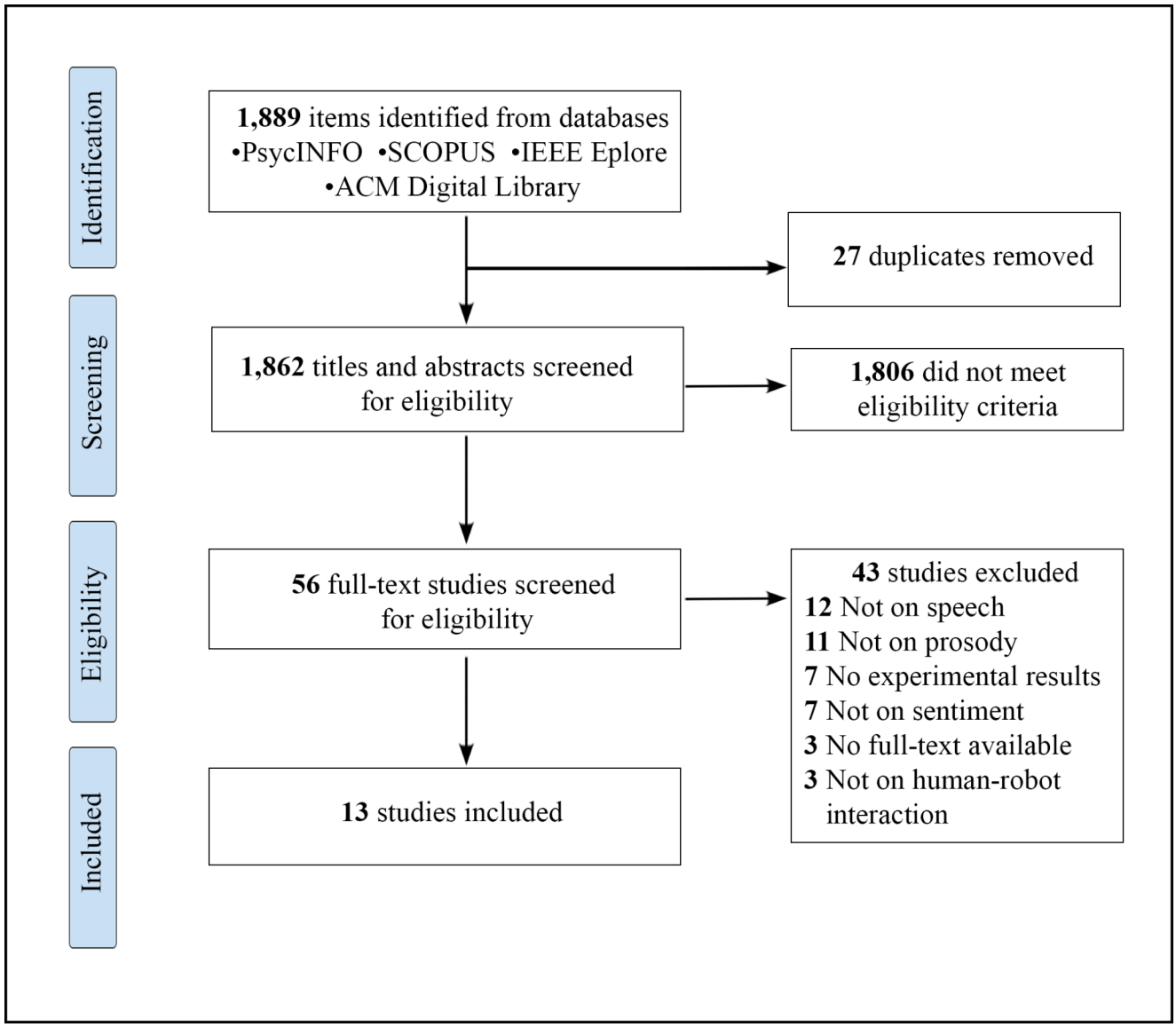

Fig. 1 PRISMA diagram showing the literature search and screening process

used Poppy, a virtual robot (embodied conversational agent) [32]. One did not provide specific details on the system used [34]. Figure 2 shows some of the robots used.

The purpose of almost all of the studies was to advance emotional speech detection and/or expression through prosody, with the ultimate purpose of implementing the system in a social robot/agent [5, 29–32, 34–39]. Thus, five studies focussed on emotion detection, four focussed on emotion expression and three focussed on both (see Table 2).

3.3 Prosodic Elements Used in the Literature

Across the literature, various different prosodic elements were used to convey or detect emotion. In six studies, the emotions included some or all of the Big Eight [5, 31, 33, 35–37]. However, Antona et al. [5] also supplemented these with emotions such as 'tired/confused' and 'focussed.' Three studies categorised emotions as positive, negative or neutral (or variations thereof) [5, 30, 36], while two used affective dimensions to determine emotion [28, 38, 39]. For example, these included activation (level of arousal), valence, power, expectation and intensity. Regardless of the approach, the emotions were sometimes determined in comparison to a baseline emotion, usually referred to as the 'neutral' or 'calm' emotion [5, 28, 33, 35, 36].

The prosodic elements from most to least common were tone (also referred to as pitch/frequency) (n = 8), loudness (also referred to as energy/volume) (n = 6), speech speed (n = 4), pauses (n = 3) and non-linguistic vocalisations (n = 1).

Table 2 Summary of the included studies

Author; year	Setting; Country	Robot/ system and purpose	Detect or express emotion	Prosodic elements	Parameters/ explanations	Evaluation method; Participants
Aly; 2015 [35]	University; France	ALICE Robot Social robot	Detect and express emotion	• Pauses/ silence • Pitch (contour and baseline) • Speech rate (ranging from lowest rate for "sadness" and the highest rate for "anger").	Sadness: baseline pitch: -4st; pitch contour: (0%,+0st)(100%,-0st); speech rate: -30%; contour features: negative-constant-negative; break time: Inter/Intra-Sentence Disgust: baseline pitch: +4st; pitch contour: (0%,-5st)(40%,-9st)(75%,-12st)(100%,-12st); speech rate: +8%; contour features: negative-exponential-negative; break time: Inter-Sentence Happiness: baseline pitch: +2st; pitch contour: (0%,+8st)(30%,+16st)(50%,+14st)(100%,+11st); speech rate: +7%; contour features: positive-parabola-positive; break time: Inter-Sentence Anger: baseline pitch: +5st; pitch contour: (0%,-18st)(50%,-14st)(75%,-10st)(100%,-14st); speech rate: +12%; contour features: negative-parabola-negative; break time: Inter-Sentence Fear: baseline pitch: +6st; pitch contour: (0%,+2st)(50%,+5st)(75%,+8st)(100%,+5st); speech rate: +7%; contour features: positive-parabola-positive; break time: Inter/Intra-Sentence	Experiment; between subjects design; survey; watching robot express and detect emotions from video recordings N = 60; university students and employees aged 20–57 years (M = 29:64, SD = 9:4).
Antona; 2019 [5]	ACE Alzheimer Center & hospital; Spain and Poland	RAMCIP (robot) Assistive social robot for older adults	Detect emotion	• Volume • Pitch • Speech rate • Pauses (extra spaces, commas and full-stops)	Emotion was divided into 3 categories: Positive, Neutral and Negative. Volume values were measured between 0 (silent)-10. Speaking rate: default value was 0. Positive values meant higher speech speed, while negative values represent slower rates of speech. The pitch value 0 was the default, while positive and negative numbers represented higher and lower pitch respectively. Neutral: Volume: 8/10, Rate: +1, Pitch: +0 Excited: Volume: 10/10, Rate: +1, Pitch: +4 Sad: Volume: 8/10, Rate: -2, Pitch: -3 Sleeping: N/A Tired/confused: same as neutral Focused: Volume: 8/10, Rate: +0, Pitch: +0	Experiment; matching phrases to emotions N = 52 adults; 42–94 years
Eyben; 2012 [28]	N/S	System for agents and robots	Detect emotion	• Loudness • F0 envelope • Prob. of voicing • Power spectrum (MFCC range 0–14) • Line Spectral Frequencies 1–8, • Log. Mel-Freq. bands 1–8	Measured 5 affective dimensions: Activation; level of arousal/active engagement vs. passiveness (i.e., boredom). Valence: pleasant (positive) vs. unpleasant (negative) emotions. Power: emotion is related to a feeling of power/control vs. weakness. Expectation: measure of unpredictability vs. familiarity. Intensity: measure of perceived emotional intensity. The most common descriptors were loudness, 0-th and 6th line spectral frequency, and MFCC 10, as well as the voicing probability. The dimensions and frequencies of descriptors were as follows: Activation: MFCC (16), log. Mel frequency bands (9), LSP frequencies (5), loudness (4), jitter (2). Expectation: MFCC (18), F0 (7), LSP frequencies (7), loudness (3), log. Mel frequency bands (2). Intensity: MFCC (11), loudness (7), LSP frequencies (6) log. Mel frequency bands (5). Power: MFCC (24), log. Mel frequency bands (3), LSP frequencies (3), F0 (2). Valence: MFCC (14), LSP frequencies (7), log. Mel frequency bands (4).	Experiment Data from SEMAINE database (summary N/S)

Table 2 (continued)

Author; year	Setting; Country	Robot/ system and purpose	Detect or express emotion	Prosodic elements	Parameters/ explanations	Evaluation method; Participants
Crumpton; 2014 [33]	University; N/S	MARY (open source speech synthesizer) used in the Survivor Buddy robot	Express emotion	• Pitch • Speed rate • Volume	Conveyed 4 emotions using semantically unpredictable text: anger, fear, happiness, and sadness. The calm vocal prosody was used as a baseline for the pitch, speech rate, and volume. For the final experiment, the following parameters were: Anger: Pitch: -50 Hz ; Pitch range: 120%; pitch contour: each word has a falling contour; speech rate: 95%; volume: 95% Calm: Pitch: unchanged; Pitch range: unchanged; pitch contour: flat; speech rate: 80%; volume: 60% Fear: Pitch: +70 Hz; Pitch range: 20%; pitch contour: rising; speech rate: 100% with random pauses between words; volume: 70% Happiness: Pitch: +50 Hz; Pitch range: 200%; pitch contour: varies between −5% and +25%; speech rate: varies between 70% and 90%; volume: 80% Sadness: Pitch: -30 Hz; Pitch range: 70%; pitch contour: falling; speech rate: 50%; volume: 40%	Experiment; survey; detecting emotion from robotic speech 52 university students (28 females and 24 males) approx. 18-19.7 years old
Hsieh; 2020 [29]	University; Japan	Pepper Robot Social robot	Express emotions	• Pitch	Combined prosodic elements (pitch) with some gestures.	1. Survey on interaction styles 2. Survey after interaction N = 31 (females 16, males 15) aged 25 ± 3
Juszkiewicz; 2014 [34]	N/S	N/S Social robot; family home	Detect emotion	• Pitch • Frequency • Energy	6 parameters were used: Intensity: instantaneous sound pressure value (measured in dB SPL). Spectrogram Pitch Mel-frequency Cepstral Coefficients: (MFCC) Harmonics to noise ratio (HNR): energy of the harmonic parts of the signal related to the energy of the noise parts. HNR was expressed in dB and computed using the autocorrelation method and the cross-correlation method. Long-Term Average Spectrum	Experiment; matching phrases to emotions N = 50
Li; 2017 [39] Li; 2019 [38]	University; N/S	ERICA robot Social robot	Detect emotion and express emotion	N/S	Considered valence and arousal	Experiment; analyse correlation between valence/ arousal and the prosodic elements in speech N = 6

Table 2 (continued)

Author; year	Setting; Country	Robot/ system and purpose	Detect or express emotion	Prosodic elements	Parameters/ explanations	Evaluation method; Participants
Rabiei; 2016 [36]	N/S; N/S	System for social robots	Detect emotion	• Pitch • Energy/ loudness • Frequency	Recognizes and classifies basic emotional states (sadness, surprise, happiness, anger, fear and disgust). The shape refers to pitch (peak, value and range) graphs. Positive emotion: low intensity, slightly right skewed (shape) Negative emotion: low intensity, slightly left skewed (shape) Negative-negative emotion: high intensity, beginning is lower & decreases sharply It uses a hybrid algorithm that combines speech graph and facial features extraction. Pitch: happiness and anger had the highest average pitch peak and sadness had the lowest. Surprise had the highest pitch value and the lowest corresponds to disgust. The fear emotion did not have a distinct peak and was similar to sadness. Intensity (loudness): surprise had the highest intensity while disgust had the lowest. Speech rate: anger and fear had the lowest speech rate (sentences pronounced with anger/fear were pronounced faster) while happiness and sadness had the highest. Happiness: Happiness had the highest average pitch peak and intensity. Surprise: had the highest pitch range and high pitch peak, right- skewed pitch contours. Anger: had the highest pitch peak, pitch values, speech rate and intensity, slightly left-skewed pitch contours	Experiment; participants repeat phrases with rising and falling intonations, intensity, speech rate and pitch movements N = 300, (150 females and 150 males, 20–48 years old)
Serban; 2017 [32]	School; France	Poppy (ECA, SEMAINE project) Social agent; Story-telling agent (ECA) for children	Express emotion	• Rhythm • Stress • Intonation /pitch	N/S	Mixed methods; Experiment; 3 conditions (no mimics, no prosody or mimics and prosody); interviews N = 50; elementary school children; 6–11 years

Table 2 (continued)

Author; year	Setting; Country	Robot/ system and purpose	Detect or express emotion	Prosodic elements	Parameters/ explanations	Evaluation method; Participants
Tsiourti; 2017 [31]	Online	Pepper and Hobbit robots Social robots	Express emotion	• Non-linguistic vocalisations(laughter, negative 'oh' and intake of breath)	Expressed 3 emotions: happiness, sadness, surprise. Non-linguistic vocalizations were synthesized using a commercial Text-to-Speech engine: laughter vocalization (happiness), negative "oh" (sadness), and a sudden, short intake of breath (surprise).	Experiment; within-subjects repeated measures design; Database of 22 videos of Hobbit and Pepper; online survey; qualitative (perception of the expression) and quantitative (recognition accuracy) responses to the expressions. N = 170
Valenti; 2020 [30]	Online	NAO (robot) (using DIARC cognitive robotic architecture) Social robots/agents	Express and detect emotion	• Pauses • Speech rate	Detected 5 states of emotional valence (strong negative, medium negative, neutral, medium positive, strong positive). Then gestured in response (to express emotion)	Experiment; online survey N/S
Yamamoto; 2018 [37]	N/S; N/S	Speech synthesiser system; VOICEROID 2 Yuduki Yukari Social robot	Detect emotion	• Volume/ loudness • Sound height (pitch) • Sound interval	Detected 3 emotions (anger, joy and sorrow). Characteristics such as volume, sound height, sound and sound interval change, across emotions were considered. For example, volume significantly increased for anger.	Observational study; differences in prosodic elements between synthesised and natural human speech N = 3

Note: MFCC (Mel Frequency Cepstral Coefficient): Coefficients that collectively make up a Mel Frequency Cepstral. These represent short-term power spectrums of sound. The Mel scale is a scale that relates the perceived frequency of a tone to the actual measured frequency.

F0: fundamental frequency.

N/S: not stated.

Fig. 2 Images showing the NAO [40], Hobbit [41] and Pepper [42, 43] robots

Changes in intonation (tone) and loudness of speech helped to express emotion. For example, a higher pitch and volume were used to express/detect a positive emotion (e.g., excitement or happiness) [5, 33, 35, 36]. Conversely, a slightly lower volume and much lower pitch expressed negative emotions (e.g., sadness) [5, 33, 35, 36]. Anger was associated with increased volume [33, 37] and a high pitch peak [36].

Speech speed and pauses were sometimes interrelated. This is logical, because adding pauses to the speech would also impact the overall speed of the speech. This was achieved by adding additional spaces in the text and adding commas and full stops to create pauses [5]. In addition, the speed also referred to the rate of speech/words spoken within a specific timeframe [5]. It was evident that a faster rate of speech correlated with the anger [33, 35, 36] excited [5] and fear [33, 36] emotions. A lower rate was associated with the sadness emotion [5, 33, 35, 36].

One study did not explain the prosodic elements and instead focussed on stress/emphasis of words and rhythm, compared to speaking in monotone [32]. Tsiourti et al. [31] also used a different approach, by synthesising commonly understood non-linguistic vocalisations in a commercial Text-To-Speech (TTS) service. These included laughter to convey happiness, a negative sounding "oh" to represent sadness and a fast sudden intake of breath to convey surprise.

3.4 Effect of the Prosodic Elements on Sentiment Detection or Expression

Most of the literature reported successful results, showing that prosodic elements are useful in helping to express or detect emotion. Some positive findings were evident in the literature on human-robot/agent interaction [29, 31, 32, 35]. For example, children smiled more and were also more responsive to questions when prosody was used in the Poppy avatar, compared to when only facial expressions were used [32]. Crumpton et al. [33] showed promising results in emotion detection, whereby participants were able to detect different emotions, above levels of chance (20%). These included anger (65.9%), calm (68.9%), fear (33.3%), sadness (49.2%) and happiness (30.3%) after adjusting some of the prosodic elements. Participants were also able to accurately identify the happy and surprised emotions from robots using non-linguistic vocalisations [31].

Mixed findings were reported in some of the literature on speech systems. The system used by Eyben et al. [28] was effective at detecting sentiment for five dimensions (activation, expectation, intensity, power and valence), outperforming standard neural networks. However, in another study the system was only effective when sentiment analysis was included in addition to prosody (increasing the correlation coefficient by 0.15, from 0.41 to 0.56) [38, 39]. This was explained due to valence conflicting with sentiment (i.e., emotional incongruence).

Negative findings were also reported. Specifically, some of the research found that several emotions were more difficult to detect by participants. These included negative emotions such as frustration, disappointment, anxiety [5], anger [35], fear, disgust [36] and sadness [31]. Aly et al. [35] explain that participants are dependent on non-verbal cues (e.g., gestures) with emotions such as anger and that the Mary TTS engine limited their ability to design a persuasive vocal pattern for this emotion. Additionally, Rabiei et al. [36] highlight that some emotions are simply more difficult for humans to identify.

4 Discussion

The most effective and commonly used prosodic elements related to emotional speech in human-computer/robot interaction were tone (n = 8), loudness (n = 6), speech speed (n = 4), pauses (n = 3) and non-linguistic vocalisations (n = 1). However, some of the literature did not specify what elements they used and instead used a lack thereof (i.e., monotone voice).

It was evident that research in this field is premature, as displayed by the small number of available studies. Additionally, this was evident in many studies focussing on the speech synthesiser systems and not yet being at the stage of implementing them in social robots/agents. However, positive findings in the literature on human-robot/agent interaction indicated a promising opportunity for implementing systems in various robots, including the popular NAO, Pepper and Hobbit robots [29–31].

It is important to note that synthesis of the findings was difficult, due to the various uses of emotion and measures of prosody. Regardless, the categorisation of emotion was consistent with that identified in the literature and often adhered to or included the Big Six [8–10], Big Eight [11], arousal and valence [12] or classification as negative, positive and neutral [9, 10, 15]. A novel finding was the addition of other emotions and dimensions. Specifically, Crumpton et al. [33] added the baseline emotion 'calm,' while Antona et al. [5] also used the 'tired/confused' and 'focussed' emotions. Some affective dimensions of determining emotion were also novel. These included the established categories of valence and arousal (also referred to as activation), but also considered power, expectation and intensity [28, 38, 39].

It was interesting that the common prosodic elements were mostly effective in helping to express or detect emotion within human-robot/agent interaction [29, 31, 32, 35],

but negative emotions were often more difficult to identify [5, 31, 35, 36]. This may be because people often rely on non-verbal cues [35]. This suggests that while prosody is important for affective computing, it is not the sole solution. Instead, this should be complemented with gestures and facial expressions (if possible), in a multimodal strategy (e.g., in [35]). Even in appearance-constrained robots, prosody can be supplemented with changes in visual appearance. For example, a study with 33 participants found that colour and motion can be combined to convey emotion [44]. Specifically, anger is best conveyed with colour, while fear can be conveyed with motion, and joy/happiness is best conveyed with a combination of colour and motion. General agreements in colour can be leveraged and employed in conjunction with speech. These included the common pairings of red to anger and yellow to joy/happiness [45, 46]. However, regardless of effectiveness, incorporating prosodic elements may further help to augment robotic voices with affective capabilities, and overcome issues with them sounding too 'robotic' as expressed in some human-robot interaction studies [5, 47].

4.1 Implications for Development and Future Research

Literature on sentiment expression and detection through prosody was fairly premature. Thus, further research is warranted, before design and development recommendations can be made. It was evident that some prosodic elements (tone, loudness and speech speed) were more often used than others (e.g., non-linguistic vocalisations), with promising results in sentiment expression and/or detection. Future research should explore the effectiveness of these specific prosodic elements in emotional speech, using larger sample sizes. Once these have been determined effective, the speech should be implemented on a robot and tested in real-life interaction scenarios. Long-term research and development should also include non-verbal parameters and verbal semantics, to determine which combination leads to the most successful expression and detection of negative sentiment.

4.2 Strengths and Limitations

The review adhered to the Arksey and O'Malley [24] method for conducting scoping reviews, and was reported in adherence with the PRISMA-ScR items [26]. Another strength of our scoping review is that the included research was limited to the last 10 years of publication, meaning that the findings represent the most recent state-of-the-art methods. This also helped us to overcome a common limitation of scoping reviews, whereby the included literature is vast [48], which may result in limited detail in the findings.

As in other scoping reviews (e.g., [49, 50]), our work was limited to literature published in English and did not include grey literature. We did also not conduct quality or bias assessments of the included literature, meaning that the included studies were of varying quality. However, this is typically not a requirement of scoping reviews [48, 50].

5 Conclusion

This scoping review of recently published literature helped to identify common prosodic elements used in human-computer interaction: tone, loudness, speech speed and pauses. Non-linguistic vocalisations and emphasis/stress were less frequently used. Future research should explore the effectiveness of commonly used prosodic elements in emotional speech, using larger sample sizes and real-life interaction scenarios. Finally, the successfulness of prosody in conveying negative sentiment may be improved with additional non-verbal parameters (e.g., motion or changes in light that represent emotion). Thus, it is essential that more work be conducted to determine how these may be combined with prosody and which combination is most effective in human-robot affective interaction.

Acknowledgements This work was supported by the Technology Innovation Program (10077553, Development of Social Robot Intelligence for Social Human–Robot Interaction of Service Robots) funded by the Ministry of Trade, Industry & Energy (MI, Korea). The authors would like to thank the CARES team and the University of Auckland Robotics Department for their support.

Authors' Contributions HSA conceived the initial idea as the principal investigator of this project. NG drafted the manuscript and designed the figures. All authors considered the results and approved the final manuscript.

Funding This work was supported by the Technology Innovation Program (10077553, Development of Social Robot Intelligence for Social Human–Robot Interaction of Service Robots) funded by the Ministry of Trade, Industry & Energy (MI, Korea).
Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflicts of Interest/Competing Interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. Picard R (1997) Affective Computing. MIT Press, USA
2. Gasteiger N, Broadbent E (2021) AI, Robotics, Medicine and Health Sciences, in The Routledge Social Science Handbook of AI, A. Elliott, Editor. Routledge, New York
3. Ochs M et al (2006) A computational model of capability-based emotion elicitation for rational agent. 1st workshop on Emotion and Computing-Current Research and Future Impact. Bremen, Germany
4. Lim J et al (2021) Subsentence Extraction from Text Using Coverage-Based Deep Learning Language Models. Sensors 21(8):2712
5. Antona M et al (2019) My robot is happy today: how older people with mild cognitive impairments understand assistive robots' affective output, in 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments. p. 416–424
6. Thompson W, Balkwill L (2006) Decoding speech prosody in five languages. Semiotica 158(1/4):407–424
7. Scherer K (1986) Vocal affect expression: A review and a model for future research. Psychol Bull 99:143–165
8. Ekman P (1992) An argument for basic emotions. Cogn Emot 6(3–4):169–200
9. Ghazi D, Inkpen D, Szpakowicz S (2010) Hierarchical versus Flat Classification of Emotions in Text. NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. LA, California, Association for Computational Linguistics, pp 140–146
10. Calix R, Javadpour L, Knapp G (2011) Detection of Affective States From Text and Speech for Real-Time Human–Computer Interaction. Hum Factors: J Hum Factors Ergon Soc 54(4):530–545
11. Plutchik R (1980) A general psychoevolutionary theory of emotion. Theories of emotion. Elsevier, pp 3–33
12. Russell J (1980) A circumplex model of affect. J Personal Soc Psychol 39(6):1161
13. Hutto C, Gilbert E (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text, in AAAI Conference on Web and Social Media.
14. Tausczik Y, Pennebaker J (2010) The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. J Lang Social Psychol 29(1):24–54
15. Aman S, Szpakowicz S (2007) In: Text S, Dialogue V, Matoušek (eds) Identifying Expressions of Emotion in Text. Springer: Berlin, Editors
16. Crystal D (1975) The English tone of voice: essays in intonation, prosody and paralanguage. The English tone of voice: essays in intonation, prosody and paralanguage. Edward Arnold, London
17. Siqueira H et al (2018) Disambiguating Affective Stimulus Associations for Robot Perception and Dialogue, in IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids). IEEE: Beijing, China
18. Provost E, Shangguan Y, Busso C (2015) Umeme: University of Michigan emotional McGurk effect data set. IEEE Transactions on Affective Computing, **6**(4): p. 395–409
19. Aguado L et al (2018) Effects of affective and emotional congruency on facial expression processing under different task demands. Acta Psychol 187:66–76
20. Paradeda R et al (2018) Would You Follow the Suggestions of a Storyteller Robot?, in 11th International Conference on Interactive Digital Storytelling. : Dublin, Ireland
21. Rodriguez I et al (2017) Adaptive emotional chatting behavior to increase the sociability of robots, in Social Robotics (ICSR), A. Kheddar, Editor. Springer, Cham
22. Anderson S et al (2008) Asking the right questions: Scoping studies in the commissioning of research on the organisation and delivery of health services. Health Res Policy Syst 6(7):1–12
23. Levac D, Colquhoun H, O'Brien KK (2010) Scoping studies: Advancing the methodology. Implement Sci 5(69):1–9
24. Arksey H, O'Malley L, Scoping studies: Towards a methodological framework.International Journal of Social Research Methodology 2005. **8**(1): p.19–32
25. Paré G et al (2015) Synthesizing information systems knowledge: A typology of literature reviews. Inform Manage 52(2):183–199
26. Tricco A et al (2018) PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 169(7):467–473
27. Moher D et al (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med 6(7):e1000097
28. Eyben F, Wöllmer M, Schuller B (2012) A multitask approach to continuous five-dimensional affect sensing in natural speech. ACM Trans Interact Intell Syst 2(1):1–29
29. Hsieh W-F, Sato-Shimokawara E, Yamaguchi T (2020) Investigation of Robot Expression Style in Human-Robot Interaction. J Robot Mechatron 32(1):224–235
30. Valenti A et al (2020) Emotion expression in a socially assistive robot for persons with Parkinson's disease, in Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments. p. 1–10
31. Tsiourti C et al (2017) Designing Emotionally Expressive Robots, in Proceedings of the 5th International Conference on Human Agent Interaction. p. 213–222
32. Şerban O et al (2017) Interactive narration with a child: impact of prosody and facial expressions, in Proceedings of the 19th ACM International Conference on Multimodal Interaction. p. 23–31
33. Crumpton J, Bethel C (2014) Conveying Emotion in Robotic Speech: Lessons Learned, in The 23rd IEEE International Symposium on Robot and Human Interactive Communication. IEEE: Edinburgh, Scotland, UK
34. Juszkiewicz L (2014) Improving Speech Emotion Recognition System for a Social Robot with Speaker Recognition, in 19th International Conference on Methods and Models in Automation and Robotics. IEEE: Miedzyzdroje, Poland
35. Aly A, Tapus A (2015) Multimodal Adapted Robot Behavior Synthesis within a Narrative Human-Robot Interaction, in International Conference on Intelligent Robots and Systems. IEEE/RSJ: Hamburg, Germany

36. Rabiei M, Gasparetto A (2016) System and method for recognizing human emotion state based on analysis of speech and facial feature extraction; Applications to Human-Robot Interaction, in nternational Conference on Robotics and Mechatronics. IEEE: Tehran, Iran
37. Yamamoto K et al (2018) Analysis of Emotional Expression by Visualization of the Human and Synthesized Speech Signal Sets, in 2018 International Workshop on Advanced Image Technology (IWAIT). IEEE: Chiang Mai, Thailand
38. Li Y et al (2019) Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. Adv Robot 33(20):1030–1041
39. Li Y, Ishi C, Ward N (2017) Emotion Recognition by Combining Prosody and Sentiment Analysis for Expressing Reactive Emotion by Humanoid Robot, in APSIPA Annual Summit and Conference. APSIPA: Malaysia
40. SoftBank Robotics, Available NAO from: https://www.softbankrobotics.com/emea/en/nao
41. TU WIen. HOBBIT - THE MUTUAL CARE ROBOT. n.d.; Available from: http://hobbit.acin.tuwien.ac.at
42. Guizzo E(2014) How Aldebaran Robotics Built its Friendly Humanoid Robot, Pepper. ; Available from: https://spectrum.ieee.org/robotics/home-robots/how-aldebaran-robotics-built-its-friendly-humanoid-robot-pepper
43. SoftBank Robotics. Pepper. n.d.; Available from: https://www.softbankrobotics.com/emea/en/pepper
44. Löffler D, Schmidt N, Tscharn R(2018) Multimodal Expression of Artificial Emotion in Social Robots Using Color, Motion and Sound, in ACM/IEEE International Conference on Human-Robot Interaction.
45. Sutton T, Altarriba J (2016) Finding the positive in all of the negative: Facilitation for color-related emotion words in a negative priming paradigm. Acta Psychol 170:84–93
46. Fugate J, Franco C (2019) What Color Is Your Anger? Assessing Color-Emotion Pairings in English Speakers. Front. Psychol
47. Gasteiger N et al (2021) Older Adults' Experiences and Perceptions of Living with Bomy, an Assistive Dailycare Robot: A Qualitative Study. Assistive Technology
48. Sucharew H, Macaluso M, Methods for Research Evidence Synthesis (2019) Scoping Rev Approach J Hosp Med 14(7):416–418
49. Gasteiger N et al (2021) Friends from the Future: A Scoping Review of Research into Robots and Computer Agents to Combat Loneliness in Older People. Clin Interv Aging 2021(16):941–971
50. Dawe J et al (2019) Can social robots help children in healthcare contexts? A scoping review. BMJ Paediatrics Open 3:e000371