Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-022-01709-y

ORIGINAL ARTICLE

Collaborative analysis of audio-visual speech synthesis with sensor measurements for regulating human–robot interaction

K. Ashok[1] · Mohd Ashraf[2] · J. Thimmia Raja[3] · Md Zair Hussain[4] · Dinesh Kumar Singh[5] · Anandakumar Haldorai[6]

Received: 19 October 2021 / Revised: 10 December 2021 / Accepted: 4 June 2022

Abstract We reside in an environment wherein robotics is used in a variety of circumstances daily. In the best-case scenario, this contact seems as natural and comfortable as human-to-human conversation. Audiovisual speech synthesis is an appropriate way of communication between a human and a robot in this case. The robot is able to communicate to its users due to audiovisual text-to-speech synthesis technology. A diverse range of approaches are conducted to synthesis audiovisual speech has been established during the previous few years. The proposed Robot Operating System (ROS) performs the collaborative analysis of audio-visual speech synthesis using sensors measurement to enable the interaction between humans and robots. Skeletal tracking, gesture identification are performed by utilizing a depth camera, as well as facial recognition utilizing an RGB camera are aspects of visual-based entities. Auditory perception is dependent on the use of a microphone array to locate sound sources. We offer a top-down hierarchy communication protocol-based integration architecture for these entities. The top layer of integration contains the message about the number of people and associated states that are changed from a number of the lower-level perceptive entity.

Keywords Audio-visual speech synthesis · Communication · Human–robot interaction · Robot operating system (ROS)

1 Introduction

The perception of the human environment is crucial in developing human–robot interaction. The robot can watch and hear activities created by humans utilizing robotic sensors, such as microphones and cameras. Various perceptual components, including recognition of face and then gesture, person tracking, and source sound identification, analyze those input signals to determine the present state (who, where, what) of every individual in the situation. Operating a robot

✉ K. Ashok
ashok@newhorizonindia.edu

Mohd Ashraf
ashraf.saifee@gmail.com

J. Thimmia Raja
thimmia@gmail.com

Md Zair Hussain
mdzairhussain@gmail.com

Dinesh Kumar Singh
dineshsingh025@gmail.com

Anandakumar Haldorai
anandakumar.psgtech@gmail.com

[1] Department of Electronics and Communication Engineering, New Horizon College of Engineering, Bengaluru 560103, India

[2] Computer Science & Engineering, School of Technology, Maulana Azad National Urdu University, Hyderabad, TS, India

[3] School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, TN, India

[4] Information Technology, School of Technology, Maulana Azad National Urdu University, Hyderabad, TS, India

[5] Department of IT, DSMNRU, Lucknow, UP, India

[6] Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu 641202, India

Published online: 09 August 2022

with specialized skills is a difficult undertaking that necessitates a wide range of knowledge from a variety of fields, involving auditory processing signal, video and audio processing, robotic planning, and then multi-modal fusion (Lane et al. 2012). For instance, combine facial recognition, tracking of audio-visual, dialog control, and speech recognition are developed in the robot for better communication with humans (Okuno et al. 2002). These modules are used to demonstrate a greeting robot in a custom framework. For such software modules to function in contemporaneous life routine, they need four computers linked in networking that are difficult for a one person to handle.

Robot Operating System has now become the most prominent and growing robotic platform (Quigley et al. 2009). It includes drivers made up of hardware devices, execution of widely used operations, and a messaging mechanism for operations to communicate with one another. It also includes numerous essential packages for building a robotic system, including navigation, perception, and SLAM (locating source and mapping). Space, voice, and gestures are included to achieve pleasant and effective human–robot interaction (HRI) (Mead and "Space 2012). Furthermore, inside the ROS ecosystem, the HRI toolkit (HRItk) is combined various elements to construct a speech synthesis interactive system. The perception nodes, on the other hand, post to a variety of themes, making it impossible to monitor data about a specific person. Thus, this toolset is quite suited to single-person HRI rather than multi-person settings. Figure 1 shows the audio-visual system model.

- *Mature field control of robot:* One that has already been widely marketed in the business world. Moreover, the techniques necessary to control human–robot contact and collaboration are still to be completely developed. Physical human–robot interaction (pHRI) (Bicchi et al. 2008), as well as collaborative robotics (CoBots), is both investigating these challenges (Colgate et al. 1996).
- *Safety:* In a collaborative analysis of robots with humans, the most essential aspect is safety. Researchers are yet in the early phases of robot safety standards, despite new efforts (e.g., robotic devices; and robots; ISO 13,482:2014, 2014). Avoiding collision (with persons or barriers) is a common security measure that necessitates high responsiveness (high bandwidth) and resilience at the control and perception layers (Khatib 1985).
- *Coexistence:* Coexistence refers to a robot's capacity to operate with humans in the same place. This com-

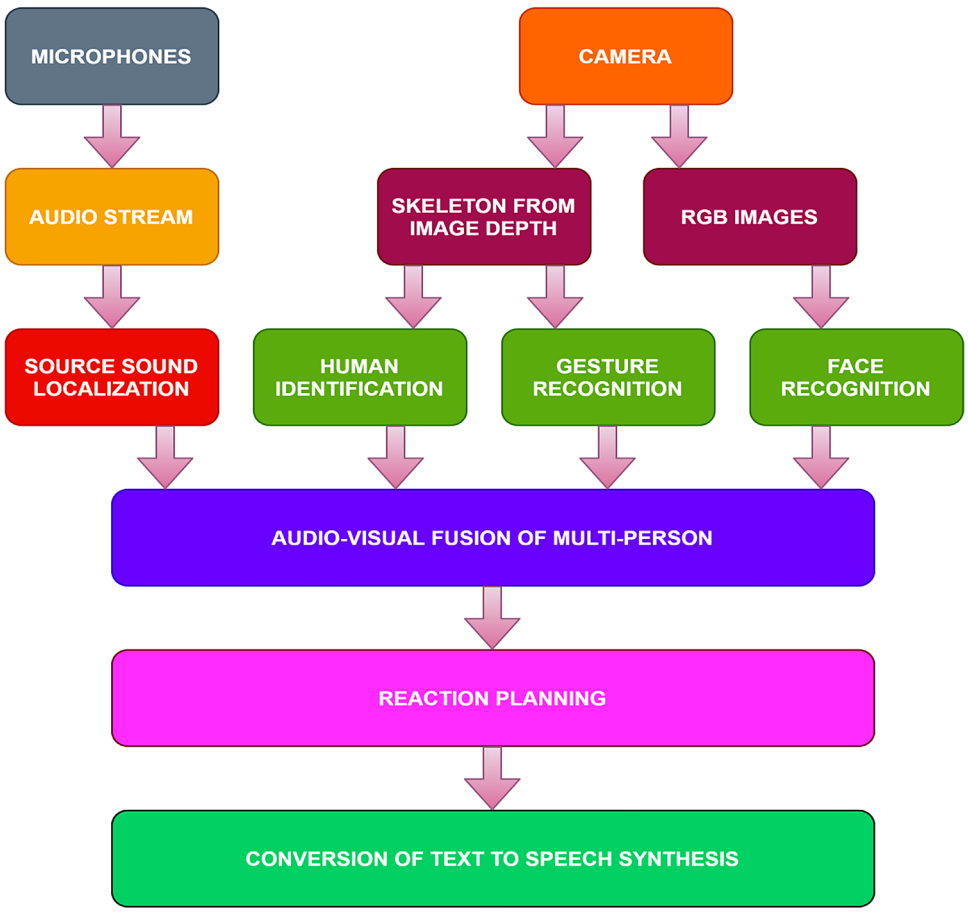

Fig. 1 Audio-Visual Integration system

prises situations wherein the robots and humans perform together on the identical task independent of contact or cooperation (e.g., medical activities in which the robot intervenes on the body of patients) (Azizian et al. 2014).

- *Collaboration:* The capacity to accomplish robot activities directly with human involvement and coordination is referred to as cooperation. Physical cooperation with explicit as well as purposeful interaction among robots and humans. The action which includes the information transformation for better human–robot interaction (HRI) includes the voice commands, body gestures, and various other modalities that have been considered as the contactless collaboration method of analysis. It is critical, particularly for the second stage, to develop mechanisms for intuitive control through human operators, who may or may not be experts.

Based on the job at hand, various combinations of sensory modalities have been used. The four fundamental robot senses consist as follows:

Vision: This covers techniques for processing and comprehending pictures in order to generate symbolic or numeric data that mimics human vision. The richness of such a sensation is unparalleled, despite the fact that picture processing is complicated as well as computationally intensive. The vision of robotics is critical for comprehending the surroundings including human intent as well as reacting appropriately.

Touch: Proprioceptive force, as well as tact, is both included in this analysis, with the former requiring physical contact directly with an exterior object. The sensation of proprioceptive force is similar to the sense of muscular force (Proske and Gandevia 2012). The robots may detect this through torque sensors else joint position errors implanted in the joints and afterward utilize both techniques to deduce and respond to human intents through controlled force (Raibert and Craig 1981; Hogan 1985; Villani and Schutter 2008). Human touch (somatosensation) is caused by the activation of the neural receptor, which is primarily found in the skin.

Audition: Binaural audition has been used to achieve sound localization in humans (i.e., ears). We may establish the source's horizontal location and elevation using auditory signals in the manner of time, level, phase discrepancies among the right and left ears (Rayleigh 1907). Artificial microphones mimic this sensation, allowing robots to find source sounds "blindly." While two microphones placed on a motorized head are commonly used in robotic hearing, alternative non-biological designs emerge such as a head equipped with a unique microphone or even an array of many Omni-directional microphones (Nakadai et al. 2006).

Distance: It is the single most important sense which humans can't directly evaluate out of the four. In the mammalian kingdom, however, several instances of echolocation may be found in whales and bats. Infrared or lidar is included within the optical sensor, ultrasonic, or capacitive sensors (Göger et al. 2010) are used by robots to detect distance. The importance of this specific "sense" in human-interaction interaction stems from the clear link between distance from barriers (in these cases, humans) and security.

2 Background

The robot may deduce motion orders such as pushing, pulling acquiring from the human through feeling force. This force detection and human movement estimate is employed depending on minimal jerk for collaborative manipulating in admittance control architecture (Maeda et al. 2001). An assistance robot reduced unintentional vibrations of a person who controlled the direction as well as welding processing speed (Suphi Erden and Maric 2011; Suphi Erden and Tomiyama 2010). The robot operation is handled with manual guidance by utilizing kinematic reduction (Markkandan et al. 2021). The publications described admission controllers for robots provided with two-arm moving a table in cooperation with a person (Perumal et al. 2021; Thangamani et al. 2020). An admission controller is used to operate a medical robotic arm (Baumeyer et al. 2015). Another frequent human–robot cooperation situation in which force feedback is important is robotic teleoperation for a detailed overview of the subject; consider (Passenberg et al. 2010). Localized force or moment metrics were used in all of these studies. Tactile sensors as well as skins (which measure the wrench throughout the robot's body; (Argall and Billard 2010) were previously been utilized for object examination (Natale and Torres-Jara 2006) or recognition (Abderrahmane et al. 2018), not for controlling. Another explanation is because these remain still in the early stages of design, which necessitates sophisticated calibration (Leonid and Jayaparvathy 2022; Lin et al. 2013), which is itself a research subject.

Li et al. 2013) are the exception, as they offered a technique that included tactile measurements. Tactile sensing had been also utilized to manage contact with the surroundings (Zhang and Chen 2000). A human–robot fabrication unit for collaborative construction of automobile joints was reported in the study (Arulaalan and Nithyanandan 2016). Through admission control, the technique (trade both touch and vision) may regulate physical interaction among robot and person, as well as among environment and robot. In hazardous scenarios, vision will take over to initiate emergency

braking. The human's positioning in relation to the robot dictated the switching criterion.

Shared control seems to be desirable in scenarios in which the human/environment and robot are in constant touch (such as collaborative object transfer). Let's start with a pioneer controller, with teleoperated pole installation by putting the loop visualization beyond the looping force. The controlling admittance distorted the standard trajectory x_r output via visual servoing in the existence of touch to get the robot location instruction x.

Employing robotic arms with a hand, presented a hybrid touch as well as a vision controller for grabbing items (Pomares et al. 2011). Touching feedback uses the fingers to grab the thing, whereas the visual input is acquired from directs an active camera (placed on the robotic tips) to monitor the object, also identify humans for avoidance. The researchers used matrix S to operate the fingers and arms individually using the appropriate sensor. A hybrid method was used to regulate an ultrasonic probe in communication with a patient's belly (Chatelain et al. 2017). The objective would be to focus on the surgeon's ultrasonic lesions.

3 Sensor-based control

3.1 Audio-based control formulation

The goal of audio-based controlling is to find the robotic movement toward the source sound. A two-dimensional binaural with two microphones arrangement of the microphone rig along with the angular velocity as the controlling input $u = \dot{\alpha}$. Interaural Time Difference (ITD), as well as Interaural Level Difference (ILD)3, are the two most prevalent techniques for quantifying error e.

The τ differential in the arriving times of the sounds on every individual microphone is used in ITD-dependent auditory servoing; τ this should be controlled to a desirable value $\tau *$. The controller may be expressed by setting $e = \tau - \tau *$ and the targeted rate as $\tau *= -\lambda(\tau - \tau *)$ (to get set-point regulating to $\tau *$). Utilizing cross-correlation conventional signals, a characteristic τ may be generated in contemporaneous. In the case of a far-field assertion:

$$e = \dot{\tau} - \dot{\tau}^* = -\left(\sqrt{(b/c)^2 - \tau^2}\right)u - \dot{\tau}^* \tag{1}$$

The sound celerity is represented as c, while the microphone baseline is denoted as b. The ITD Jacobian's scalar form is represented as $J_\tau = -\sqrt{(b/c)^2 - \tau^2}$ based on (1). The motion which reduces e to the smallest value is:

$$u = -\lambda J_\tau^{-1}(\tau - \tau^*) \tag{2}$$

That is locally specified for $\alpha \in (0, \pi)$ to guarantee that $|J_\tau| \neq 0$.

The difference in strength among the right and left signals ρ is used in ILD-dependent aural servoing. This may be calculated as $\rho = \frac{E_l}{E_r}$ in a time frame of size N, wherein $E_{l,r} = \sum_{n=0}^{N} \gamma l,r[n]^2$ denotes the sound energy of the signals whereas γl, $r[n]$ denotes the intensity at iteration n. $e = \dot{\rho} - \dot{\rho}^*$ along with $\dot{\rho}^* - \lambda(\rho - \rho^*)$ is being used to control ρ to a desired ρ^*. Considering spherical propagation as well as a signal that changes gradually:

$$e = \dot{\rho} - \dot{\rho}^* = \frac{y_s(\rho + 1)b}{L_r^2}u - \dot{\rho}^* \tag{3}$$

Here, y_s denotes the forward coordinate of the source sound in the movable audiovisual frame, whereas L_r is the length among the source as well as the right microphone. The formation of ILD Jacobian scalar representation is given as $J_\rho = y_s(\rho + 1)b/L_r^2$. The motion which reduces e to the smallest value is:

$$u = -\lambda J_\rho^{-1}(\rho - \rho^*) \tag{4}$$

J_ρ^{-1} refers to sources that are positioned in advance of the rig. Unlike ITD-servoing, the location of the source supply (i.e., y_s as well as L_r) should be specified or approximated in this case. Whereas the techniques mentioned only manage the angular velocity of rigs ($u = \dot{\alpha}$), Magassouba expanded both to manage the mobile system's 2D translations.

Because of the nature of such a sensation, audio-based controls are typically employed in contact-free applications, to augment other senses (such as force and length) with sound, or to create natural human–robot interactions. Audio-based controlling is presently an undeveloped experimental topic with a lot of promise for human–robot collaboration, such as speaker tracking. Some have phrased the issue diversely from the mentioned publications, which followed closely the structure. The developed linear model represents the relationship among a robotic head's pan movement and then the intensity differential among the two microphones. The resultant acquired from controllers has been significantly simpler rather than (2) and (4). However, because their working range was narrower, they were less robust than their highly analytical rivals. Figure 2 shows the voice remote control system.

3.2 Sound source localization

The sound source localization (SSL) unit uses a microphone array to identify a sound occurrence then estimate the location of the source sound. Because it has an

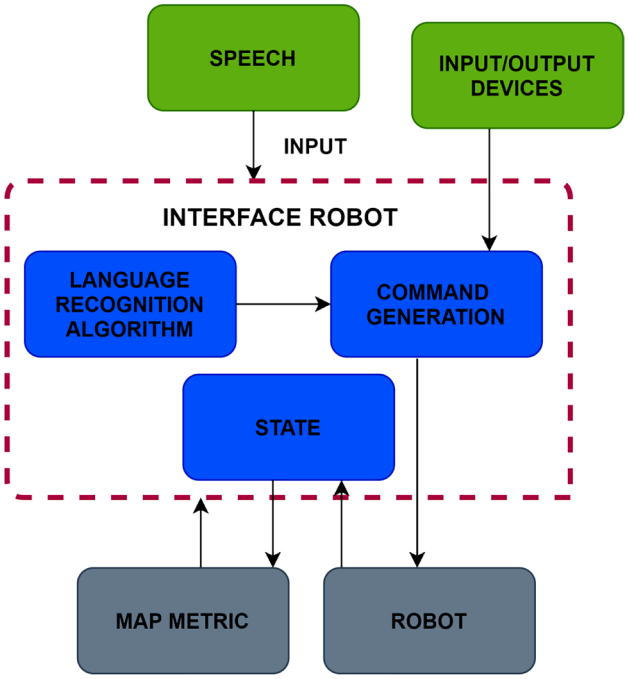

Fig. 2 Voice remote control system

inbuilt 4 microphones array, the Kinect sensor might be utilized for SSL. Moreover, this array seems linear and therefore, could only locate sound sources on 1⁄2 of the plane (180°), causing front-back confusion. Thus, the four microphones provided from the microphone array are being mounted on the robotic head portion. The HARK library has been used to develop the SSL unit on the Flow designer middleware. To determine if a frame comprises just surrounding sound or a destination source sound, a speech activity recognition relying on short energy is being used. The Phase Difference of Arrival (PDOA) method is then used to determine the direction from the sound frame. Such estimates are collected for multiple successive frames and afterward grouped to determine the sound direction of the event (azimuthal angle).

3.3 Visual servoing formulation

The employment of vision to regulate robotic mobility is referred to as visual servoing. The cameras might be set in the workstation or installed on a movable component of the robotic. “Eye-to-hand” and “eye-in-hand” visual servoing are the terms used to describe these distinct setups. The error e is specified in terms of certain picture characteristics, indicated by s, that must be controlled to a desirable configuration s^* (s is equivalent to x in the previous inverse kinematic description). The visual mistake is as follows:

$$e = \dot{s} - \dot{s}^* \tag{5}$$

When s is specified in image space, then the visual servoing methods are termed depending on image, and then whether s is described in 3-dimensional operational space, they are termed position-based. The method depending on image (in its eye-in-hand modalities) is just shortly mentioned herein since the approach based on position entails the task projection from the picture to an operating space to achieve x.

The most basic controller depending on the image utilizes $s = [X, Y]^T$, where X, as well as Y, are the picture pixel coordinates, to create u, which directs s towards reference $s^* = [X^*, Y^*]^T$. It is accomplished by describing e as follows:

$$\dot{s} - \dot{s}^* = \begin{bmatrix} \dot{X} - \dot{X}^* \\ \dot{Y} - \dot{Y}^* \end{bmatrix}, \text{with } \dot{s}^* = -\lambda \begin{bmatrix} X - X^* \\ Y - Y^* \end{bmatrix} \tag{6}$$

While considering the camera’s 6-dimensional velocity as $u = v_c$ control input, The Jacobian matrix[2] of an image connecting $\left[\dot{X}, \dot{Y}\right]^T$ as well as u are:

$$J_v = \begin{bmatrix} \frac{-1}{\varsigma} & 0 & \frac{X}{\varsigma} & XY & -1 - X^2 & Y \\ 0 & \frac{-1}{\varsigma} & \frac{Y}{\varsigma} & 1 + Y^2 & -XY & -X \end{bmatrix} \tag{7}$$

Here, represents the point’s depth in relation to the camera when there are no restrictions, it is:

$$v_c = -J_v^+ \lambda \begin{bmatrix} X - X^* \\ Y - Y^* \end{bmatrix} \tag{8}$$

3.4 Skeleton tracking and gesture recognition using depth images

Regarding skeletal tracking, this study used to package 1 of openni_tracker. This employs image depth to follow the human skeleton in real-time. The feet, hips, torso, knees, hands, elbows, neck, shoulders, and head are among the joints. This is feasible to identify human motions via buffering those joints. Elbows and hands, for instance, can be tracked to detect "hand waving."

Assuming $P_{joint}(t) = \left\{\vec{P}_{joint}(t); \vec{P}_{joint}(t-1); \ldots\ldots; \vec{P}_{joint}(t-W)\right\}$ denote the sum of each and every joint $\vec{P}_{joint}(t) = \left[P^x_{joint}(t), P^y_{joint}(t), P^z_{joint}(t)\right]$ locations across W successive frames. When the elbow is at a stationary correspondingly the hand moves exclusively in a horizontal plane, then the hand is waved:

$$g(t) = \begin{cases} 1 \; if \sigma^x_{hand}(t) \;\; gt; TH_{hand}, \sigma^x_{elbow}(t) \;\; lt; \varepsilon \\ 0 \; otherwise \end{cases} \tag{9}$$

Here, joint $\sigma_{joint}(t)$ represents the standard deviation of the $P_{joint}(t)$ collected set, TH_{hand} represents a threshold of hand motion, but ε is a threshold of elbow motion. W is fixed to 30 frames in this study, and the thresholds are $TH_{hand} = 0.1$ as well as $\varepsilon = 0.01$, which were determined from a preliminary study.

Moreover, the head position at a specific height is represented as $P^{z}_{head}(t)$ the head determines two additional stages of a person such as "standing" and "seated."

4 Performance evaluation

Every speech synthesis segment (i.e., per keyframe) is given a goal vector of parametric values in this proposed audio-visual speech model. To mix the desired values across time, functional dominating overlapping temporal are utilized. The dominance processes are exponent functions of two negative components, in which one rising whereas the other decreases. For every sample articulator and phoneme control characteristic, the peak height and the pace dominantly fall and rises are free characteristics that may be modified. The ROS enables the human–robot interaction with the interpolated face model parameter among keyframes is defined by the dominant features of the speech sections. Figure 3 shows the parameter tracking, dominant features, and target values for the human–robot interaction via the facial modeling for the efficient ROS. There are a variety of methods for assessing the efficacy of an audiovisual speech synthesis which may be classed as objective, perceptive, or subjective assessment procedures. These three important features are also assessed.

We examined three alternative approaches for acquiring denoised characteristics with regard to MFCCs as well as LMFB audio characteristics in the first study. Figure 3, as well as Figs. 4, and 5, illustrates speech recognition and synthesis levels for mel-frequency cepstral coefficients (MFCCs) as well as log mel-scale fiterbank (LMFB) characteristics tested with various SNRs values for sound inputs for the proposed systematic approach. These findings show that MFCCs outperformed LMFB in most cases. When contrasted to the original input, the audio characteristic obtained by combining successive different images with sensor measurements has a better noise resilience. MFCC and LMFB characteristic features are evaluated for the ROS systematic approach for better human–robot interaction for various component values namely 8, 16, and 32.

5 Conclusion

In this paper, the audio-visual speech synthesis is obtained using the proposed robot operating system (ROS) for efficient human–robot interaction (HRI) in multi-person settings. The suggested system includes sound source localization, face identification, and recognition, gesture recognition, which are all necessary aspects for HRI. The robots are provided with an RGB camera as well as microphone array is used to illustrate this architecture. This system is considered a foundation system for HRI because it is made up of numerous open-source apps. Moreover, certain components are retained for assessing methods using facial recognition as a ground-truth source. While vision, as well as touch, is the most common modalities on collaborative robots today,

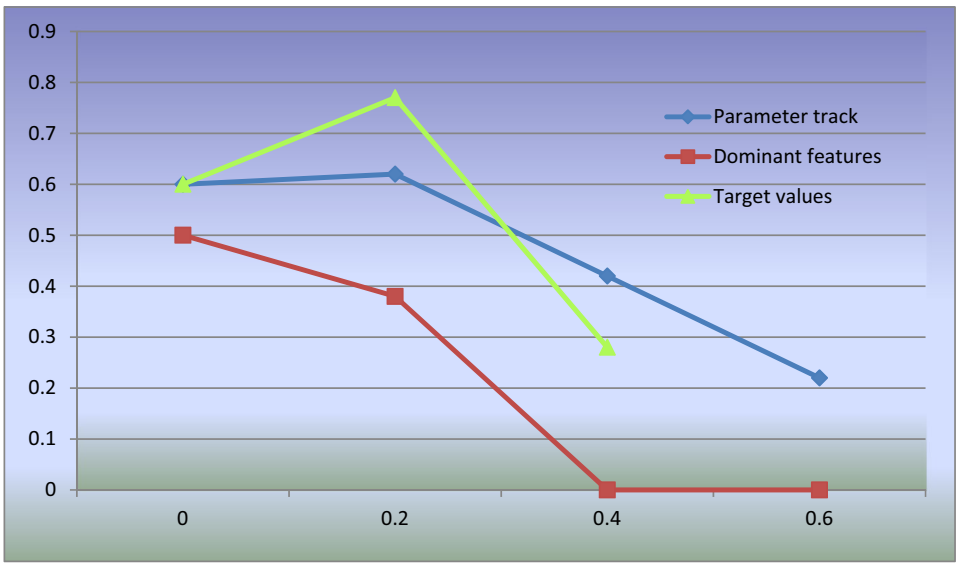

Fig. 3 Visual coarticulation modeling for ROS

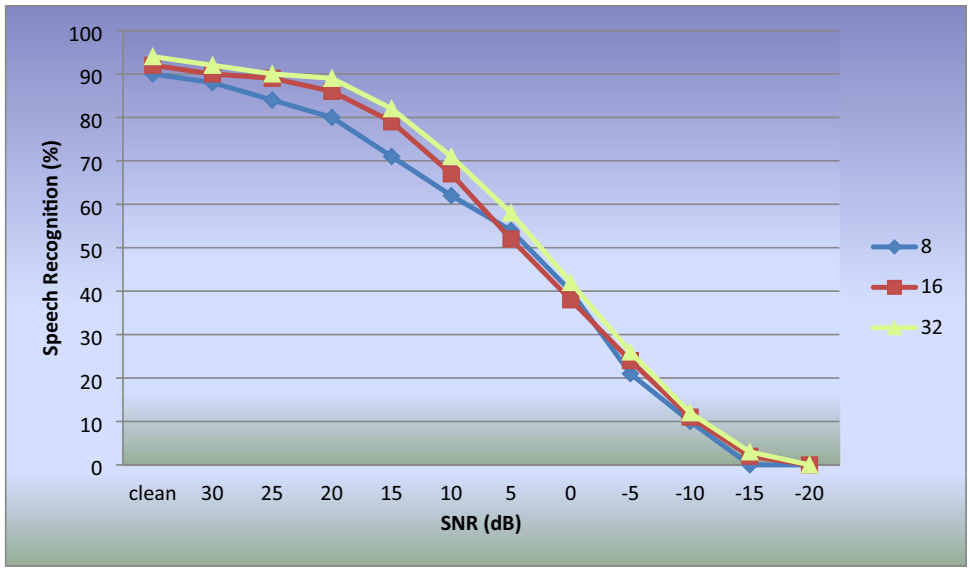

Fig. 4 MFCC speech recognition

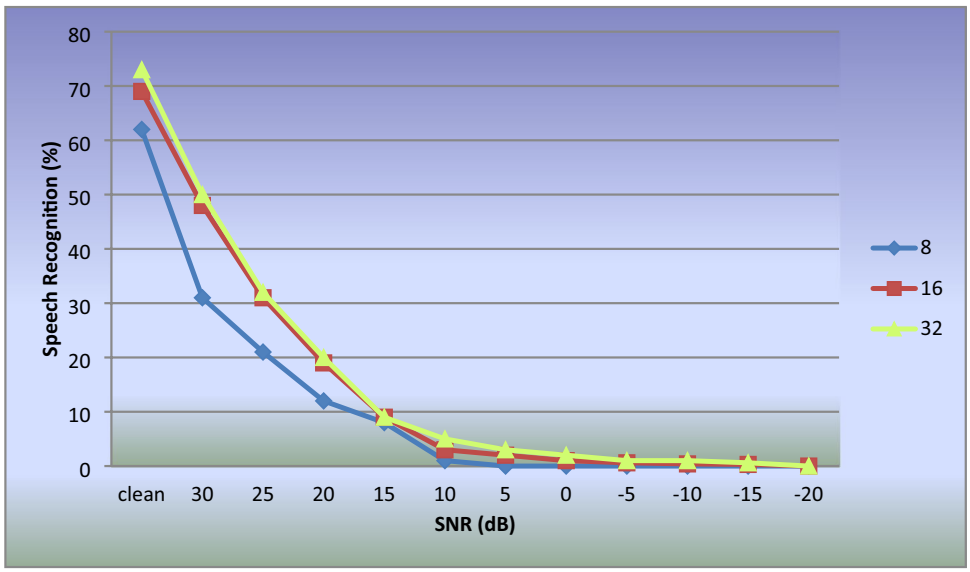

Fig. 5 LMFB speech recognition

the introduction of inexpensive, accurate, and easy-to-integrate distance, tactile, and auditory sensors opens up exciting possibilities for the future.

Funding The authors received no specific funding for this study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The manuscript has not been submitted to more than one journal for simultaneous consideration. The manuscript has not been published previously. The Research not involved human participants and/or animals.

References

Abderrahmane Z, Ganesh G, Crosnier A, Cherubini A (2018) Haptic zero-shot learning: recognition of objects never touched before. Robot Auton Syst 105:11–25. https://doi.org/10.1016/j.robot.2018.03.002

Argall BD, Billard AG (2010) A survey of tactile human-robot interactions. Robot Auton Syst 58:1159–1176. https://doi.org/10.1016/j.robot.2010.07.002

Arulaalan M, Nithyanandan L (2016) Dual band triangular microstrip antenna for WLAN/WiMAX applications. Int J Commun Antenna Propag (Ire.C.A.P.) 6(3):132–137

Azizian M, Khoshnam M, Najmaei N, Patel RV (2014) Visual Servoing in medical robotics: a survey. Part I: endoscopic and direct vision imaging-techniques and applications. Int J Med Robot 10:263–274. https://doi.org/10.1002/rcs.1531

Baumeyer J, Vieyres P, Miossec S, Novales C, Poisson G, Pinault S (2015) Robotic co-manipulation with 6 DOF admittance control: application to patient positioning in proton-therapy. In: IEEE Int. Work. on Advanced Robotics and its Social Impacts, pp 1–6. https://doi.org/10.1109/ARSO.2015.7428220

Bicchi A, Peshkin M, Colgate J (2008) Safety for physical human-robot interaction. Springer Handbook of Robotics. https://doi.org/10.1007/978-3-540-30301-5_58

Chatelain P, Krupa A, Navab N (2017) Confidence-driven control of an ultrasound probe. IEEE Trans Robot 33:1410–1424. https://doi.org/10.1109/TRO.2017.2723618

Colgate J, Wannasuphoprasit W, Peshkin M (1996) Cobots: robots for collaboration with human operators. Proc ASME Dynamic Syst Control Div 58:433–439

Göger D, Blankertz M, Wörn H (2010) A tactile proximity sensor. IEEE Sensors pp 589–594

Hogan N (1985) Impedance control: an approach to manipulation: parts I-III. ASME J Dyn Syst Measure Control 107:1–24. https://doi.org/10.1115/1.3140701

Khatib O (1985) Real-time obstacle avoidance for manipulators and mobile robots. In: IEEE Int. Conf. on Robotics and Automation, ICRA. https://doi.org/10.1109/ROBOT.1985.1087247

Lane I, Prasad V, Sinha G, Umuhoza A, Luo S, Chandrashekar A, Raux A (2012) HRItk: The human-robot interaction toolkit rapid development of speech-centric interactive systems in ros. In: Proceedings of the ACL-STDS, pp 41–44

Leonid TT, Jayaparvathy R (2022) Classification of elephant sounds using parallel convolutional neural network. Intell Autom Soft Comput 32(3):1415–1426

Li Q, Schürman C, Haschke R, Ritter H (2013) A control framework for tactile servoing. Robotics Science and Systems (RSS). https://doi.org/10.15607/RSS.2013.IX.045.

Lin CH, Fishel JA, Loeb GE (2013) Estimating point of contact, force and torque in a biomimetic tactile sensor with deformable skin. Technical report, SynTouch LLC

Maeda Y, Hara T, and Arai T (2001) Human-robot cooperative manipulation with motion estimation. In: IEEE/RSJ Int. Conf. on Robots and Intelligent Systems, IROS, Vol. 4, pp 2240–2245. https://doi.org/10.1109/IROS.2001.976403

Markkandan S, Narayanan L, Robert Theivadas J, Suresh P (2021) Edge based interpolation with refinement algorithm using EDGE strength filter for digital camera images. Turk J Physiother Rehabil 32(2):981–993

Mead R (2012) Space, speech, and gesture in human-robot interaction. In: Proceedings of international conference on multimodal interaction (ICMI), pp 333–336

Nakadai K, Nakajima H, Murase M, Kaijiri S, Yamada K, Nakamura T, et al. (2006) Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays. In: IEEE Int. Conf. on Acoustics Speech and Signal Processing. https://doi.org/10.1109/ICASSP.2006.1661122

Natale L, and Torres-Jara E (2006) A sensitive approach to grasping. In: Proceedings of the 6th Int. Workshop on Epigenetic Robotics

Okuno HG, Nakadai K, Kitano H (2002) Social interaction of humanoid robot based on audio-visual tracking. In: Hendtlass T, Ali M (eds) Developments in applied artificial intelligence. Springer, Berlin Heidelberg, pp 725–735

Passenberg C, Peer A, Buss M (2010) A survey of environment-operator and task-adapted controllers for teleoperation systems. Mechatronics 20:787–801. https://doi.org/10.1016/j.mechatronics.2010.04.005

Perumal S, Tabassum M, Narayana G, Suresh P, Chakraborty C, Mohanan S, Basit Z, Quasim MT (2021) ANN base novel approach to detect node failure in wireless sensor network. Comput Mater Continua (TechScience) 69(2):1447–1462

Pomares J, Perea I, Garcia GJ, Jara CA, Corrales JA, Torres F (2011) A multi-sensorial hybrid control for robotic manipulation in human-robot workspaces. Sensors 11:9839–9862. https://doi.org/10.3390/s111009839

Proske U, Gandevia SC (2012) The proprioceptive senses: their roles in signaling body shape, body position and movement, and muscle force. Physiol Rev 92:1651–1697. https://doi.org/10.1152/physrev.00048.2011

Quigley M, Gerkey B, Conley K, Faust J, Foote T, Leibs J, Berger E, Wheeler R, Ng A (2009) ROS: an open source robot operating system. In: Proceedings of the open-source software workshop of the international conference on robotics and automation (ICRA)

Raibert MH, Craig JJ (1981) Hybrid position/force control of manipulators. ASME J Dyn Syst Meas Control. https://doi.org/10.1115/1.3139652

Rayleigh L (1907) On our perception of sound direction. Lond Edinburgh Dublin Philos Mag J Sci 13:214–232. https://doi.org/10.1080/14786440709463595

Suphi Erden M, Maric B (2011) Assisting manual welding with robot. Robot Comput Integr Manufact 27:818–828. https://doi.org/10.1016/j.rcim.2011.01.003

Suphi Erden M, Tomiyama T (2010) Human intent detection and physically interactive control of a robot without force sensors. IEEE Trans Robot 26:370–382. https://doi.org/10.1109/TRO.2010.2040202

Thangamani M, Ganthimathi M, Sridhar SR, Akila M, Keerthana R, Ramesh PS (2020) Detecting coronavirus contact using internet of things. Int J Pervasive Comput Commun 16(5):447–456. https://doi.org/10.1108/IJPCC-07-2020-0074

Villani L, De Schutter J (2008) Force control. In: Siciliano B, Khatib O (eds) Springer handbook of robotics. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 161–185. https://doi.org/10.1007/978-3-540-30301-5_8

Zhang H, Chen NN (2000) Control of contact via tactile sensing. IEEE Trans Robot Autom 16:482–495. https://doi.org/10.1109/70.880799