Voice Command Recognition for Explicit Intent Elicitation in Collaborative Object Transportation Tasks: a ROS-based Implementation

J. E. Domínguez-Vidal
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Barcelona, Spain
jdominguez@iri.upc.edu

Alberto Sanfeliu
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Barcelona, Spain
alberto.sanfeliu@upc.edu

ABSTRACT

Voice command recognition remains relatively unexplored in robotics, with limited insight into user acceptance and real-world performance. In this work we try to address this by offering multiple voice command recognition models encapsulated in a single publicly available ROS node ready to be used by the robotics practitioner. We tested its actual performance with 10 volunteers of different nationalities whose first spoken language is not English. The obtained accuracy in these tests varies between 93.14% and 95.63% depending on the number of considered commands and model size. Finally, we conducted a user study with 23 new volunteers performing a human-robot collaborative transport task to test whether humans are willing to use this type of system despite having a non-negligible delay and failure rate. In addition to improvements in parameters such as comfort and trust in the robot, 86.9% of the volunteers chose this system over a technically more robust one.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; *Empirical studies in HCI*.

KEYWORDS

Human-Robot Interaction, Intention Understanding, Voice Recognition, Human-centered Studies, Natural Language Processing

ACM Reference Format:

J. E. Domínguez-Vidal and Alberto Sanfeliu. 2024. Voice Command Recognition for Explicit Intent Elicitation in Collaborative Object Transportation Tasks: a ROS-based Implementation. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640749

1 INTRODUCTION

Since its appearance more than 100 years ago, the term "robot" has evolved along with its capabilities and tasks. We started with isolated automatons that performed routine and repetitive tasks. Gradually these tasks became more complex but their execution was still performed in isolation, without interaction with any human. Subsequently, robots began to leave the industrial environments and began to interact with us. First dodging us as if we were obstacles, then with small punctual interactions and now starting to perform fully collaborative tasks in which an almost constant interaction between human and robot is required.

This evolution is largely due to the increasing ability to detect and interpret human intent with which we have been endowing the robot. Starting with simple ways of modeling the motion of the passers-by [18] to using increasingly elaborate motion predictors [15, 16, 21]. Starting from more classical architectures [14, 39] to modern Deep Learning-based models [24, 29, 38, 43].

While the previous examples are cases in which an attempt is made to understand the human's intention implicitly, that is, by inferring it from their actions [9, 10]; in recent years an attempt has been made to obtain the human's intention explicitly, i.e., by trying to communicate directly with them either by using user interfaces of different types [8, 12, 13] or by using more natural means of communication such as gestures [7, 30] or natural language [23, 28]. Within the latter group, the use of voice commands is a simpler first approach that, although there are Deep Learning models with high success rates [27, 33], they have rarely been used in robotics. Even less has been tested whether humans are willing to use them or how the price to pay (delay, system with non-negligible failure rate) affects them.

In this work, we take several Deep Learning models already designed (no contribution in this aspect) and encapsulate them in a ROS (Robot Operating System) [32] node whose repository we make publicly available together with its installation instructions, being this our first contribution. This done, we check the actual success rate of each model with people of different nationalities and, therefore, different accents from those present in the training dataset [40] typically used in the literature to train all these models. In this way, the user of our node can know what is the expected performance as well as the inference time being this our second contribution. Finally, we performed a round of real experiments with 23 volunteers in which we used a collaborative transportation task to compare two systems for eliciting explicit human intention: a voice command recognition system and a button-based system. In this way, we test whether the human is willing to accept a more natural communication system but with a higher delay and failure rate, an assumption that we usually take for granted but that has not been tested, this being our third contribution.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03…$15.00
https://doi.org/10.1145/3610978.3640749

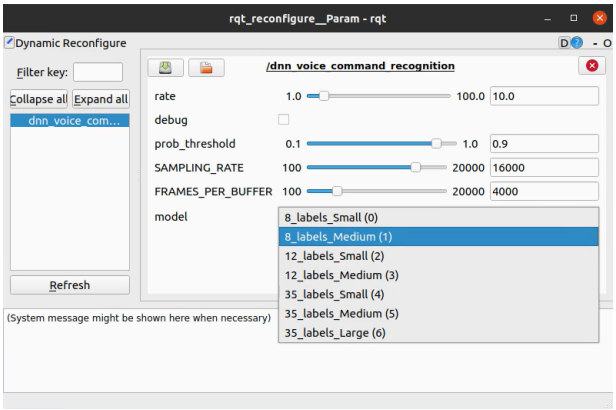

Figure 1: User interface of our voice command recognition node using ROS tool 'rqt_reconfigure'. It shows the params that the user can tune to modify its performance.

In the remainder of the article, Section 2 presents the work related to this article and Section 3 shows the capabilities of our node. Section 4 includes the real-world accuracy of each model tested with real humans as well as its inference time and shows the results of our round of experiments with the two systems. Finally, Section 5 presents the conclusions and future work.

2 RELATED WORK

Speech command recognition is a task that has been attempted for decades using increasingly elaborate techniques from rudimentary Hidden Markov Models (HMM) [6] to systems based on Mel-Frequency Cepstral Coefficients (MFCCs) and Vector Quantization (VQ) [35] or Dynamic Time Warping (DTW) [4]. However, it was not until the proliferation of Artificial Neural Networks (ANNs) that satisfactory results for finite sets of commands began to be obtained [20, 27, 33, 42]. In turn, this proliferation of Deep Learning based models has fostered the emergence of datasets containing lists of typical commands being [40] the most widely used due to containing more than 100.000 samples divided into 35 commands. In this work, we will be based on [20, 42] to implement the different models that will be trained and validated first on the mentioned dataset and later with users from our research center.

Applied to robotics, the first attempts to communicate verbally with a robot by means of voice commands in combination with gestures [34] suffered from ambiguity problems that limited the possibilities of these systems to control the robot. Subsequent works [26, 31] have made use of more elaborate techniques that can transmit simple movement commands to a mobile robot. More recent articles seek either to achieve better command detection systems [36] or to use them in tasks where more complex interactions are required, such as surgery [44] or industrial environments [19]. However, all of these works have in common that they assume that humans want to use such systems despite their failure rate simply because it comes more naturally to them. In this work we conduct a user study to test whether this assumption holds.

As for the collaborative transport of objects, this task is usually performed in close proximity between the human and the robot so that a fast response system is needed in order to do not annoy or bother the human. It consists of moving an object at a short distance so that the robot only needs to move its end-effector or at a longer distance so that the robot must also move its platform. It is common to perform this task by using controllers [1, 5, 37, 41] so that the robot adapts in the best possible way to the trajectory desired by the human. There are also works in which they use some kind of predictor of the trajectory [3], the velocity profile [2], or even the force to be exerted by the human [11] to know their implicit intention. Less common is to find works where the human can explicitly communicate with the robot using, for example, gestures to tell the robot where they want to take the object [25]. To the best of our knowledge, this work is the first one in which voice commands are used in this specific use-case.

3 VOICE COMMAND RECOGNITION THROUGH DEEP LEARNING MODELS

The dataset mentioned in the previous section [40] is typically used to train Deep Learning models in two variants. First, using all samples in the dataset to detect up to 35 commands. Secondly, using a reduced version of it with only the 12 commands from this list: ['background_noise', 'down', 'go', 'left', 'no', 'off', 'on', 'right', 'stop', 'unknown', 'up', 'yes']. The 'background_noise' category is obtained from audio samples in which no command is mentioned and the 'unknown' category using samples of the other commands present in the dataset. The usual procedure in the literature consists of processing these samples, with a duration not exceeding one second, to obtain their spectrogram as an image so that Convolutional Neural Networks (CNNs) can be used.

In this work we consider three versions of the dataset: the full dataset with all 35 labels, a reduced dataset with 12 labels and an extra reduced dataset with only 8 labels (discarding the options 'background_noise', 'off', 'on' and 'unknown'). For each of the datasets we train between 2 and 3 models (in the case of the full dataset) with different sizes in terms of number of parameters and based on CNNs and ResNets [17] as these are the workhorses used in [20, 42]. Thus, the 12- and 35-label versions can be compared with the State of the Art and the 8-label version is available in case the reader needs more precision using only the most basic commands. In total, we obtain 7 models that offer different options when choosing between accuracy and computational load.

All these models are encapsulated in a single ROS node which we put publicly available[1] and whose user interface is shown in Fig. 1. Through this, the user can choose the rate at which the node is executed, whether or not to display debug messages, the minimum probability with which the node must recognize a command for it to be published or the specific model the user wants to use.

As for its internal operation, this is controlled by the parameters SAMPLING_RATE and FRAMES_PER_BUFFER. The node tries to create audio chunks of 1 second duration (this is the duration of the samples in the dataset) so SAMPLING_RATE determines the frequency in *Hz* at which the audio signal is sampled and the number of samples that each chunk must contain in order to be sent to the selected model to predict the voice command. On the other hand,

[1]GitHub repository: https://github.com/JEDominguezVidal/dnn_voice_command_recognition

Table 1: Performance obtained with each of the tested models

	Model	Accuracy [%]		Inference Time [ms] (min./avg./max.)	
		In testset split	In real experiments	RTX 2060 Mobile	RTX 3060 Mobile
8 labels	Small (0.71 M)	97.75	95.00	6.8 / 17.0 / 33.0	5.2 / 8.6 / 15.1
	Medium (1.76 M)	98.14	95.63	9.9 / 18.5 / 34.6	5.4 / 9.0 / 14.5
12 labels	Small (0.71 M)	96.12	93.75	6.8 / 16.8 / 27.6	5.1 / 8.5 / 13.8
	Medium (1.76 M)	97.24	94.75	10.3 / 18.7 / 33.4	5.2 / 8.9 / 14.3
35 labels	Small (0.72 M)	95.19	93.14	6.9 / 16.9 / 30.6	5.2 / 8.7 / 16.4
	Medium (2.95 M)	95.63	93.71	10.7 / 23.5 / 35.1	5.5 / 9.5 / 18.6
	Large (11.2 M)	97.13	94.29	24.7 / 36.9 / 63.4	12.3 / 16.3 / 27.8

FRAMES_PER_BUFFER determines the number of old audio samples that the node must replace with new samples to create a new chunk. In other words, between one audio chunk and the next one there is an overlap of SAMPLING_RATE - FRAMES_PER_BUFFER samples. This is done to ensure that if a voice command is said out loud right between two audio chunks, this command does not go undetected. A higher overlap offers a better guarantee that this does not happen but at the cost of increased computational cost.

Additionally, the node repository also includes Google Colab Notebooks with all the necessary code so that each of the models can be retrained if desired or their internal structure modified.

4 RESULTS

Now that we have this ROS node with all the trained models we can, first, test its actual performance in a known environment and with people not present in the dataset. Secondly, we can use this node in a specific task where the human can use voice commands to guide the humanoid robot IVO [22]. In this way, we can compare this system with one that is less natural but eliminates the problems of delay and failure rate associated with command recognition and check if the human is indeed willing to use this type of system.

All the experiments reported in this document have been performed under the approval of the ethics committee of the Universitat Politècnica de Catalunya (UPC) in accordance with all the relevant guidelines and regulations (ID: 2023.05).

4.1 Performance in real experiments

The original articles on which the models trained in this work are based indicate the level of accuracy that each model can obtain in the dataset used to train them. Here we report the actual performance obtained with each version of each model both in the testset split and by using this model later in real tests with volunteers.

For this purpose, 10 people (age: $\mu = 29.50$, $\sigma = 4.89$; self-reported spoken English skills from 1 to 7: $\mu = 4.88$, $\sigma = 0.83$) are recruited from our research center with different nationalities and, therefore, different spoken accents. This allows us to obtain a more accurate idea of its real-world performance in environments where English is not the first language of the system users.

In this case, the nationalities of the people who have tested the system would be as follows: 5 Spanish, 1 Chinese, 1 French, 1 Iranian, 1 Italian, 1 Mexican. They each read aloud the full list of 35 commands once, pausing for several seconds after each command. Subsequently, they read the reduced list of 10 commands (the list of 12 labels with out the 'background_noise' and 'unknown' cases) three times, also pausing for several seconds after each command. At the same time, a measurement of the ambient noise level is taken before each new volunteer starts reading to know the conditions in which the system is being tested, obtaining a mean value of 40.9±2.3 dBA (see Fig. 2 for an example of the setup used). In this way, we obtain 1 sample per participant for the list of 35 commands and 4 samples per participant for the list of 10 commands. With this, we generate our own datasets of 35 labels (350 samples) and 10 labels (400 samples) with which we can check the real performance of each model. Table 1 summarizes the result obtained.

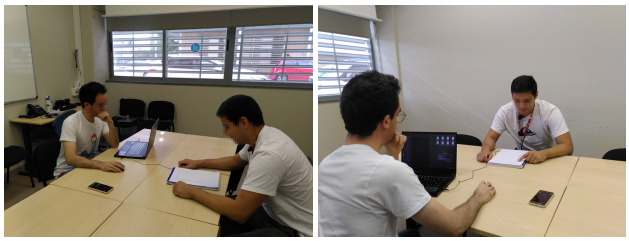

Figure 2: Setup used to obtain our own testsets. Quite room with a researcher on one side of the table recording the audio samples and volunteer on the other side reading each list of commands. Measurement of ambient noise in the middle before start reading each list of commands.

The accuracy obtained by the different models varies between 95.19% and 98.14% in the testset split of the original dataset and between 93.14% and 95.63% in our testsets obtained with real users of different nationalities. Overall, there is a drop in performance of between 2.05% and 2.75%. Table 1 also reports the inference time obtained for each model using two different graphics cards to get an idea of the required computational load. It is worth mentioning that the main delay component of this type of system consists of forming the audio chunk before sending it to the inference model.

4.2 Acceptability Study

To test whether the human really wants to use such systems we use a collaborative transportation task as a use case. In it, the pair transports an object through a scenario with multiple walls and obstacles so that there are multiple routes to the goal. Thus, the human must tell the robot which route to follow at every intersection.

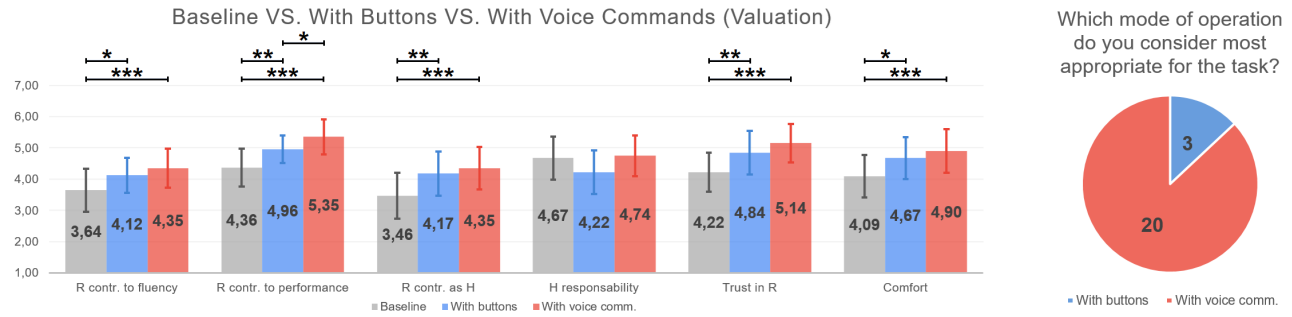

Figure 3: Assessment of the main aspects involved in the interaction. *Left:* **Comparison among the baseline experiment (without buttons or voice commands) in gray, experiment with buttons in the handle in blue and voice commands in red. Valuation from 1 (very low) to 7 (very high). Statistical significance marked with *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. 'R' means Robot and 'H' means Human.** *Right:* **Election made by the 23 volunteers with respect to which system they prefer for the task at hand.**

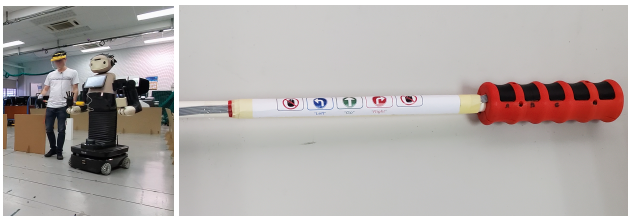

Figure 4: Use case for acceptability study. *Left:* **Human and robot collaboratively transport an aluminium bar through a maze.** *Right:* **Transported bar with buttons in the handle and meaning of each button annotated next to it.**

We recruit 23 new volunteers (age: $\mu = 27.36$, $\sigma = 4.87$) who perform three experiments each. In the first, they can only communicate with the robot by exerting force on the transported object, which is measured and interpreted by the robot using a force sensor on its wrist. In the second experiment, in addition to using their force, they can also communicate explicitly with the robot by using three buttons on the handle of the object telling it which route to follow. In the third, instead of buttons, they can use our voice recognition system (8-labels Medium model) and the commands 'Go', 'Left', and 'Right' with the same purpose[2] (see Fig. 4). To avoid statistical distortions, the order of the second and third experiments is randomized. After each experiment, volunteers fill out a questionnaire rating multiple aspects of the interaction. All variables analyzed are normally distributed according to the Shapiro-Wilk test unless otherwise stated. Fig. 3 shows the results obtained.

Applying to each variable an ANOVA test to check if there are statistically significant differences (according to the criterion of $p < 0.05$) and then a *post hoc* Tukey's HSD (Honest Significant Difference) test in case there are, it can be observed that the use of the command recognition system outperforms in all the aspects analyzed the system with buttons to explicitlly communicate with the robot despite the fact that the latter does not present appreciable delays and can not make any inference mistake. We highlight the increases with respect to the base experiment in the contribution of the Robot to fluency ($F(2, 66)$=7.66; with buttons: p=0.032; with voice: $p < 0.001$), trust in the Robot ($F(2, 66)$=12.19; with buttons: p=0.005; with voice: $p < 0.001$) and comfort ($F(2, 66)$=8.67; with buttons: p=0.014; with voice: $p < 0.001$).

Additionally, volunteers were asked at the end of the experiments about which system they consider more appropriate for the task founding that 86.9% of them prefer the system with voice command recognition. Therefore, it can be affirmed that the human does accept to use this type of systems, although an in-depth analysis would be necessary to know all the reasons behind this choice.

[2]Experiments example: https://youtu.be/R7NlOYfpl5c

5 CONCLUSIONS AND FUTURE WORK

In this work we deliver a series of State of the Art models for voice command recognition encapsulated in a ROS node so that the robotics practitioner can use this tool without the need to be familiar with Deep Learning techniques. We have also shown the performance of each model, not only in dataset but in controlled environments with volunteers of different nationalities for whom English is not their first language. We believe that this aspect is useful to know how these models work in the real world. Finally, we have conducted a user study in which we found that humans prefer to use this system to communicate explicitly with the robot over others that may be technically more robust but less human-like.

In the experiments performed with the robot, the delay problems and higher failure rate are due to the use of a Bluetooth microphone and the existence of ambient noise caused by the robot wheels. This can be minimized by connecting the microphone directly to the robot and using multiple microphones so that noise cancellation can be performed. As future work, we plan to use this system in other tasks such as handover or collaborative search. We consider that this work can serve as a basis for future studies in which the robot is provided with full natural language processing capabilities.

ACKNOWLEDGMENTS

Work supported under the European project CANOPIES (H2020-ICT-2020-2-101016906) and by JST Moonshot R & D Grant Number: JPMJMS2011-85. The first author acknowledges Spanish FPU grant with ref. FPU19/06582.

REFERENCES

[1] Don Joven Agravante, Andrea Cherubini, Antoine Bussy, Pierre Gergondet, and Abderrahmane Kheddar. 2014. Collaborative human-humanoid carrying using vision and haptic sensing. In *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 607–612.

[2] Ali Al-Yacoub, YC Zhao, William Eaton, Yee Mey Goh, and Niels Lohse. 2021. Improving human robot collaboration through Force/Torque based learning for object manipulation. *Robotics and Computer-Integrated Manufacturing* 69 (2021), 102111.

[3] Konstantinos I Alevizos, Charalampos P Bechlioulis, and Kostas J Kyriakopoulos. 2020. Physical human–robot cooperation based on robust motion intention estimation. *Robotica* 38, 10 (2020), 1842–1866.

[4] Anjali Bala, Abhijeet Kumar, and Nidhika Birla. 2010. Voice command recognition system based on MFCC and DTW. *International Journal of Engineering Science and Technology* 2, 12 (2010), 7335–7342.

[5] Antoine Bussy, Pierre Gergondet, Abderrahmane Kheddar, François Keith, and André Crosnier. 2012. Proactive behavior of a humanoid robot in a haptic transportation task with a human partner. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 962–967.

[6] Etienne Cornu, Nicolas Destrez, Alain Dufaux, Hamid Sheikhzadeh, and Robert Brennan. 2002. An ultra low power, ultra miniature voice command system based on hidden markov models. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, IV–3800.

[7] Xavier Cucurull and Anaís Garrell. 2023. Continual Learning of Hand Gestures for Human-Robot Interaction. *arXiv preprint arXiv:2304.06319* (2023).

[8] Marc Dalmasso, J.E. Domínguez-Vidal, Iván J. Torres-Rodríguez, Anaís Garrell, and Alberto Sanfeliu. 2023. Shared Task Representation for Human-Robot Collaborative Navigation: The Collaborative Search Case. *International Journal of Social Robotics* (2023). https://doi.org/10.1007/s12369-023-01067-0

[9] J.E. Domínguez-Vidal, Nicolás Rodríguez, and Alberto Sanfeliu. 2024. Perception-Intention-Action Cycle in Human-Robot Collaborative Tasks: the Collaborative Lightweight Object Transportation Use-Case. *International Journal of Social Robotics* (2024), to appear.

[10] J. E. Domínguez-Vidal, Nicolás Rodríguez, and Alberto Sanfeliu. 2023. Perception-Intention-Action Cycle as a Human Acceptable Way for Improving Human-Robot Collaborative Tasks. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. ACM/IEEE, 567–571. https://doi.org/10.1145/3568294.3580149

[11] J. E. Domínguez-Vidal and Alberto Sanfeliu. 2023. Improving Human-Robot Interaction Effectiveness in Human-Robot Collaborative Object Transportation using Force Prediction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7839–7845.

[12] J. E. Domínguez-Vidal and Alberto Sanfeliu. 2023. Inference VS. Explicitness. Do We Really Need the Perfect Predictor? The Human-Robot Collaborative Object Transportation Case. In *32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1866–1871.

[13] J. E. Domínguez-Vidal, Iván J. Torres-Rodríguez, Anaís Garrell, and Alberto Sanfeliu. 2021. User-friendly smartphone interface to share knowledge in human-robot collaborative search tasks. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 913–918.

[14] Gonzalo Ferrer and Alberto Sanfeliu. 2014. Bayesian human motion intentionality prediction in urban environments. *Pattern Recognition Letters* 44 (2014), 134–140.

[15] Gonzalo Ferrer and Alberto Sanfeliu. 2014. Proactive kinodynamic planning using the extended social force model and human motion prediction in urban environments. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1730–1735.

[16] Oscar Gil, Anaís Garrell, and Alberto Sanfeliu. 2021. Social Robot Navigation Tasks: Combining Machine Learning Techniques and Social Force Model. *Sensors* 21, 21 (2021). https://doi.org/10.3390/s21217087

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Dirk Helbing and Peter Molnar. 1995. Social force model for pedestrian dynamics. *Physical review E* 51, 5 (1995), 4282.

[19] Wojciech Kaczmarek, Jarosław Panasiuk, Szymon Borys, and Patryk Banach. 2020. Industrial robot control by means of gestures and voice commands in off-line and on-line mode. *Sensors* 20, 21 (2020), 6358.

[20] Byeonggeun Kim, Simyung Chang, Jinkyu Lee, and Dooyong Sung. 2021. Broadcasted residual learning for efficient keyword spotting. *arXiv preprint arXiv:2106.04140* (2021).

[21] Javier Laplaza, Albert Pumarola, Francesc Moreno-Noguer, and Alberto Sanfeliu. 2021. Attention deep learning based model for predicting the 3D Human Body Pose using the Robot Human Handover Phases. *30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2021), 161–166. https://doi.org/10.1109/RO-MAN50785.2021.9515402

[22] Javier Laplaza, Nicolás Rodríguez, J. E. Domínguez-Vidal, Fernando Herrero, Sergi Hernández, Alejandro López, Alberto Sanfeliu, and Anaís Garrell. 2022. IVO Robot: A New Social Robot for Human-Robot Collaboration. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 860–864.

[23] Zhihao Li, Yishan Mu, Zhenglong Sun, Sifan Song, Jionglong Su, and Jiaming Zhang. 2021. Intention understanding in human–robot interaction based on visual-NLP semantics. *Frontiers in Neurorobotics* 14 (2021), 610139.

[24] Zitong Liu, Quan Liu, Wenjun Xu, Zhihao Liu, Zude Zhou, and Jie Chen. 2019. Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing. *Procedia CIRP* 83 (2019), 272–278.

[25] Viktor Lorentz, Manuel Weiss, Kristian Hildebrand, and Ivo Boblan. 2023. Pointing Gestures for Human-Robot Interaction with the Humanoid Robot Digit. In *32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, to appear.

[26] Xiaoling Lv, Minglu Zhang, and Hui Li. 2008. Robot control based on voice command. In *2008 IEEE International Conference on Automation and Logistics*. IEEE, 2490–2494.

[27] Somshubra Majumdar and Boris Ginsburg. 2020. Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition. *arXiv preprint arXiv:2004.08531* (2020).

[28] Jinpeng Mi, Hongzhuo Liang, Nikolaos Katsakis, Song Tang, Qingdu Li, Changshui Zhang, and Jianwei Zhang. 2020. Intention-related natural language grounding via object affordance detection and intention semantic extraction. *Frontiers in Neurorobotics* 14 (2020), 26.

[29] Jae Sung Park, Chonhyon Park, and Dinesh Manocha. 2019. I-planner: Intention-aware motion planning using learning-based human motion prediction. *The International Journal of Robotics Research* 38, 1 (2019), 23–39.

[30] Marc Peral, Alberto Sanfeliu, and Anaís Garrell. 2022. Efficient Hand Gesture Recognition for Human-Robot Interaction. *IEEE Robotics and Automation Letters* 7, 4 (2022), 10272–10279.

[31] Snejana Pleshkova, Zahari Zahariev, and Alexander Bekiarski. 2018. Development of speech recognition algorithm and labview model for voice command control of mobille robot motio. In *2018 international conference on high technology for sustainable development (HiTech)*. IEEE, 1–4.

[32] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng, et al. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. Kobe, Japan, 5.

[33] Nicolae-Catalin Ristea, Radu Tudor Ionescu, and Fahad Shahbaz Khan. 2022. Septr: Separable transformer for audio spectrogram processing. *arXiv preprint arXiv:2203.09581* (2022).

[34] Oliver Rogalla, Markus Ehrenmann, R Zollner, Regine Becher, and Rüdiger Dillmann. 2002. Using gesture and speech control for commanding a robot assistant. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*. IEEE, 454–459.

[35] Mahdi Shaneh and Azizollah Taheri. 2009. Voice command recognition system based on MFCC and VQ algorithms. *World Academy of Science, Engineering and Technology* 57 (2009), 534–538.

[36] Yuuki Tada, Yoshinobu Hagiwara, Hiroki Tanaka, and Tadahiro Taniguchi. 2020. Robust understanding of robot-directed speech commands using sequence to sequence with noise injection. *Frontiers in Robotics and AI* 6 (2020), 144.

[37] Sonny Tarbouriech, Benjamin Navarro, Philippe Fraisse, André Crosnier, Andrea Cherubini, and Damien Sallé. 2019. Admittance control for collaborative dual-arm manipulation. In *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE, 198–204.

[38] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[39] Peter Trautman, Jeremy Ma, Richard M. Murray, and Andreas Krause. 2013. Robot navigation in dense human crowds: the case for cooperation. In *2013 IEEE International Conference on Robotics and Automation (ICRA)*. 2153–2160. https://doi.org/10.1109/ICRA.2013.6630866

[40] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).

[41] Xinbo Yu, Bin Li, Wei He, Yanghe Feng, Long Cheng, and Carlos Silvestre. 2021. Adaptive-constrained impedance control for human–robot co-transportation. *IEEE transactions on cybernetics* 52, 12 (2021), 13237–13249.

[42] Yundong Zhang, Naveen Suda, Liangzhen Lai, and Vikas Chandra. 2017. Hello edge: Keyword spotting on microcontrollers. *arXiv preprint arXiv:1711.07128* (2017).

[43] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. 2022. Spatio-temporal gating-adjacency gcn for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6447–6456.

[44] Kateryna Zinchenko, Chien-Yu Wu, and Kai-Tai Song. 2016. A study on speech recognition control for a surgical robot. *IEEE Transactions on Industrial Informatics* 13, 2 (2016), 607–615.