Analisis *Word Error Rate* dan Waktu Respons pada Sistem *Question-Answering* di Robot *Humanoid*

Donny Prasetya Hutagalung[1], Eko Rudiawan Jamzuri[2]

{donnyhutagalung7@gmail.com[1], ekorudiawan@polibatam.ac.id[2]}

Department of Electrical Engineering, Politeknik Negeri Batam, Batam, Indonesia[1,2]

Abstract. Penelitian ini mengusulkan sistem *Question-Answering* (QA) melalui media suara. *Automatic Speech Recognition* (ASR) dikembangkan menggunakan model VOSK untuk mengenali lima pertanyaan yang diucapkan oleh penutur secara acak dan QA dikembangkan menggunakan model RoBERTa. Hasil pengenalan suara di ASR kemudian memberikan jawaban sesuai dengan pertanyaan yang diucapkan. Dari pengujian yang dilakukan pada lima penutur secara acak dengan 125 kali percobaan, diperoleh nilai *Word Error Rate* (WER) sebesar 0.187. Sementara itu, sistem QA memiliki waktu respons dengan rata-rata sebesar 464,04 *milliseconds*. Hasil ini menunjukkan bahwa masih terdapat beberapa kesalahan pada ASR yang mempengaruhi kinerja sistem QA secara keseluruhan dan waktu respons sistem memberikan pengalaman yang cukup responsif. Hasil penelitian ini memberikan kontribusi pada penelitian dan pengembangan sistem QA, khususnya pada robot *humanoid* yang masih belum banyak diteliti.

Keywords: *Automatic Speech Recognition*, *Question-Answering*, *Word Error Rate*, Waktu respons.

1 Introduction

Salah satu bidang penelitian yang sedang berkembang pesat, terutama seiring berkembangnya era Industri 4.0 adalah *Human-Robot Interaction* (HRI). Di era Industri 4.0, teknologi automasi dan robotik harus ramah terhadap manusia dan lingkungan sekitarnya [1], [2]. Hal ini membawa cara berinteraksi antara manusia dengan robot memiliki variasi yang beragam. Salah satu cara interaksi yang dapat digunakan adalah melalui media suara [3]. Dalam hal ini, robot harus memiliki keterampilan komunikasi, termasuk mendengarkan, berbicara, dan berinteraksi [4]. Kemampuan robot mendengarkan dan memahami apa yang diucapkan manusia dapat disebut dengan *Automatic Speech Recognition* (ASR). ASR merupakan teknologi yang digunakan untuk mengenali kata dari ucapan manusia serta mengubahnya menjadi teks yang dapat dipahami oleh komputer.

Interaksi manusia dengan robot semakin mendekati pengalaman antar manusia. Salah satu contoh adalah robot yang dilengkapi dengan kemampuan *Question-Answering* (QA) [5], yang memungkinkan robot berkomunikasi dengan manusia menggunakan media suara. Pendekatan QA pertama-tama mentransfer konten lisan ke dalam transkrip ke dalam transkrip teks melalui

ASR, kemudian menggunakan beberapa metode yang efektif seperti pencocokan kesamaan [6], pencarian informasi [7] untuk memprediksi jawaban yang diberikan transkripsi ASR [8].

Meskipun saat ini QA menjadi topik penelitian di bidang HRI yang sedang berkembang pesat, namun kenyataannya tantangan dalam ASR menjadi penting dalam konteks implementasi sistem *Question-Answering* pada robot. Kendala utama terletak pada kemampuan ASR dalam mengenali kata dari ucapan manusia menjadi teks dengan akurasi tinggi. Dalam konteks QA, ketidakakuratan ASR dapat menyebabkan kesalahan interpretasi terhadap pertanyaan yang diajukan oleh pengguna. Misalnya, kata-kata seperti, *to*, *two*, dan *too* dalam sering diprediksi salah. Kesalahan ini dapat berakibat fatal jika sistem QA diimplementasikan pada robot, seperti robot pelayan medis [9], [10], [11], [12], robot pelayan restoran [13], [14], [15], dan penerapan di bidang lainnya.

Ketidakakuratan ASR tidak hanya mempengaruhi pemahaman pertanyaan tetapi juga dapat merugikan kinerja sistem QA. Jawaban yang dihasilkan oleh robot menjadi tidak sesuai dengan pertanyaan yang sebenarnya diajukan. Di satu sisi, saat ini kebutuhan terhadap robot mulai meningkat. Kebutuhan yang meningkat terhadap robot dalam berbagai peran menunjukkan pentingnya pengembangan teknologi ASR agar dapat memberikan pengalaman interaksi melalui media suara. ASR tidak hanya berkontribusi pada peningkatan akurasi pengenalan suara tetapi juga mendukung respons cepat pada sistem QA. Waktu respons juga merupakan faktor untuk memastikan kinerja yang memuaskan dalam sebuah sistem. Ketika sistem dapat dengan cepat merespons, mentranskripsikan ucapan, dan menghasilkan jawaban, pengguna mendapatkan keuntungan dari umpan balik yang lebih cepat dan pengalaman yang lebih lancar. Respons tinggi dalam mengubah sinyal suara menjadi teks memungkinkan pengguna berkomunikasi dengan sistem secara lebih efektif [16].

Dalam konteks pengembangan sistem *Question-Answering* di robot *humanoid* yang menggunakan aktivitas pengenalan suara bawaan dari model VOSK, penelitian ini berfokus pada analisis *Word Error Rate* (WER) dan waktu respons sistem. Sebagai langkah awal, pada penelitian ini, sistem dirancang untuk dapat menghasilkan jawaban sesuai dengan pertanyaan yang diajukan oleh penutur *non-native* dalam bahasa Inggris. Lima pertanyaan yang menjadi *input* dari jawaban adalah "*where is the immigration office?*", "*where is the mayor's office?*", "*where is the airport?*", "*where is the nearest gas station?*", dan "*where is the nearest hospital?*".

Susunan bagian yang tersisa dari artikel ini adalah sebagai berikut. Kita dapat memulai dengan mengulas tentang metode yang terdiri dari empat aspek: *automatic speech recognition*, *speech recognition module*, *question-answering module*, dan *evalution method*. Hasil penelitian dan diskusi disajikan di bagian tiga. Akhirnya, bagian empat memberikan kesimpulan singkat dari penelitian ini.

2 Materials and methods

Bagian ini menguraikan metode yang digunakan dalam penelitian ini. Bagian ini pertama-tama akan menjelaskan speech recognition. Kemudian akan menjelaskan *models* yang diperlukan untuk pemrosesan pertanyaan.

2.1 Speech Recognition

Automatic Speech Recognition merupakan kemampuan sebuah sistem memproses ucapan manusia ke dalam format tertulis [17]. Pada **Fig. *1***, proses ASR diawali dengan interaksi manusia yang berbicara. Ucapan manusia direkam sebagai sinyal suara dalam bentuk gelombang suara atau waveform. Kemudian, sinyal suara melewati serangkaian tahapan pemrosesan, termasuk filtering dan ekstraksi fitur. Selanjutnya, ASR menganalisis fitur-fitur dan hasil pengenalan suara ini ditampilkan dalam bentuk teks. Dalam sistem ASR, pemrosesan phoneme sangat penting, di mana ucapan manusia dibagi menjadi satuan-satuan suara terkecil yang membedakan arti setiap kata dalam suatu bahasa, contohnya: "*tree*" : /t/ - /r/ - /i/ [18].

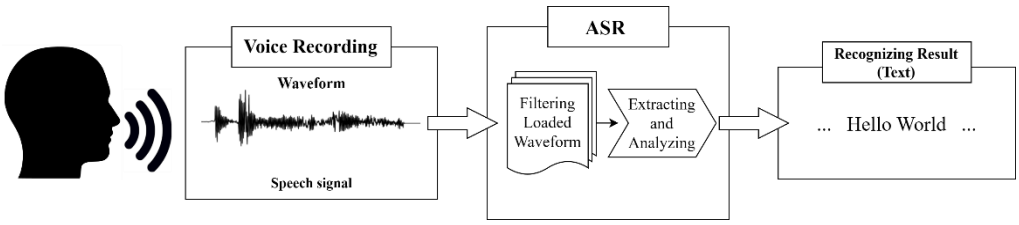

Fig. 1. *Automatic Speech Recognition Process.*

2.2 *Models*

Sistem ini dirancang menggunakan dua model yaitu *speech recognition* model untuk mengenali suara dan *question-answering* model untuk merespons pertanyaan. Kedua model ini digunakan agar sistem dapat mengenali pertanyaan dan menghasilkan jawaban yang tepat.

2.2.1 *Speech Recognition Model* (VOSK API)

Speech Recognition Model ini adalah perpustakaan baru dari CMUSphink dan disebut VOSK. VOSK adalah *speech recognition module* yang bersifat *open-source* dengan fungsi lebih lanjut. Menurut repositori Github, VOSK telah dan terus dikembangkan sejak 2019. Versi API saat ini adalah 0.3.45 [19]. API ini digunakan untuk pengenalan suara yang memiliki tingkat keakuratan dalam menampilkan data *speech-to-text*. Model dalam API ini merupakan algoritma yang digunakan untuk melakukan pengenalan suara yang dibangun dengan menggunakan teknologi *deep learning* untuk mempelajari pola dalam suara dan mengenali kata-kata yang diucapkan.

2.2.2 *Question-Answering Model* (RoBERTa)

Robustly Optimized BERT *pre-training Approach* (RoBERTa) adalah sebuah model *Natural Language Processing* (NLP) dan merupakan versi modifikasi dari model NLP yang populer, *Bidirectional Encoder Representations from Transformers* (BERT). Model ini lebih seperti sebuah pendekatan yang lebih baik untuk melatih dan mengoptimalkan BERT [20]. RoBERTa dilatih pada korpus besar data bahasa Inggris dengan pengawasan mandiri. Model ini menyamarkan 15% kata dalam *input* secara acak, lalu menjalankan seluruh kalimat yang disamarkan melalui model dan harus memprediksi kata-kata yang disamarkan [21].

Graphical User Interface (GUI) [22] dirancang untuk mengenali suara, menghitung WER, menghitung waktu respons, dan memberikan jawaban terhadap pertanyaan yang diberikan seperti yang ditunjukkan pada **Fig. *2***. Dalam sistem ini, suara akan dikenali setelah menekan tombol start. Setelah suara dikenali, maka sistem akan menghasilkan WER dari pertanyaan yang tersebut dan waktu respons yang diperlukan untuk menghasilkan jawaban.

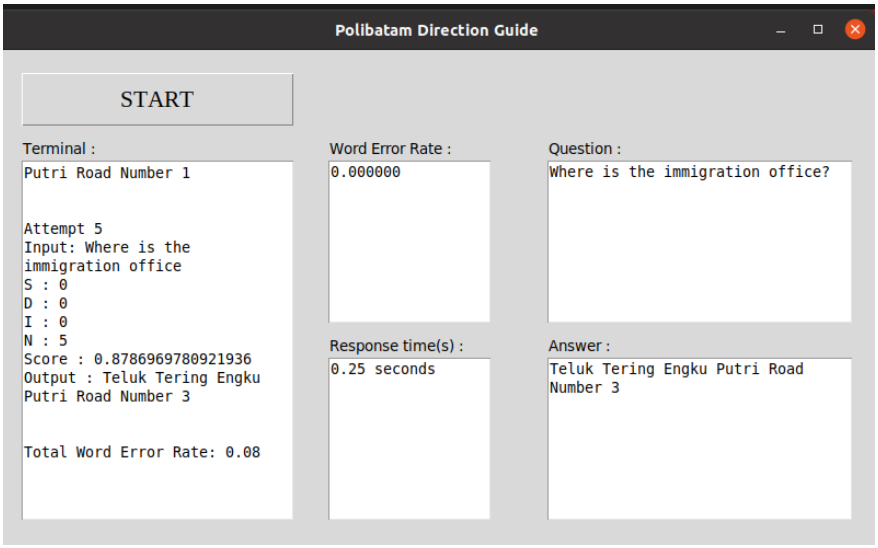

Fig. 2. *Question-Answering System Implementation.*

3 Results and discussion

Bagian ini membahas hasil penelitian yang telah diperoleh. Pembahasan diawali dengan penjelasan mengenai metode evaluasi yang dilakukan dalam penelitian ini. Selain itu, terdapat penjelasan bagaimana melakukan pengambilan data untuk memverifikasi hasil penelitian ini. Kemudian, penulis menjelaskan hasil yang diperoleh dari beberapa percobaan. Hasil tersebut dijelaskan pada sub-bab 3.1 dan sub-bab 3.2 di bawah ini.

3.1 *Evaluation methods*

Penulis menggunakan dua metode evaluasi untuk menguji metode yang diusulkan. *Word Error Rate* (WER) dan waktu respons adalah dua metode pengujian yang digunakan.

Pengujian WER dilakukan untuk mengetahui akurasi dalam merespons memahami ucapan dari penutur *non-native* berbahasa inggris. Sementara itu, waktu respons menjadi parameter penting dalam mengukur kinerja sistem QA secara *real-time*.

3.1.1 *Word Error Rate*

WER adalah salah satu metrik evaluasi yang umum digunakan dalam bidang ASR. WER digunakan untuk mengukur seberapa dekat hasil transkripsi teks yang dihasilkan oleh ASR dengan teks referensi sebenarnya [23].

$$WER = \frac{S + D + I}{N} \tag{1}$$

Persamaan (1) merepresentasikan rumus sederhana untuk menghitung WER. Di mana S, D, dan I adalah jumlah kata yang diganti, dihapus, dan disisipkan. Sedangkan N adalah total kata yang diucapkan [24].

S (*substitution*) adalah ketika sistem menangkap sebuah kata, tetapi kata tersebut adalah kata yang salah. Sebagai contoh, sistem dapat menangkap "*The cat is sleeping*" dan bukannya, "*The at is sleeping*", D (*deletions*) adalah kata yang tidak disertakan oleh sistem. Seperti, "*The cat sleeping*", I (*insertions*) adalah ketika sistem memasukkan kata yang tidak diucapkan, "*The little cat is sleeping*", dan N (*total number of words spoken*) adalah banyaknya kata dalam transkripsi referensi.

3.1.2 *Response Time*

Waktu respons dari sebuah sistem memiliki dampak signifikan terhadap penggunaannya. Waktu respons yang diperlukan sistem memengaruhi kualitas interaksi pengguna, keakuratan, dan efektivitas komunikasi. Waktu respons yang optimal merupakan aspek penting dalam meningkatkan kinerja sistem [25].

$$\Delta t = t - t_0 \tag{2}$$

Waktu respons dari sebuah sistem adalah waktu yang diperlukan saat sistem memulai mengenali suara hingga sistem menghasilkan jawaban [26]. Persamaan (2) merepresentasikan rumus sederhana untuk menghitung waktu respons yang diperlukan sistem ini. Di mana t_0 adalah waktu awal ketika sistem akan mengenali suara, t adalah akhir eksekusi ketika sistem menghasilkan jawaban, dan Δt adalah waktu respons yang diperlukan sistem untuk mengenali suara dan menghasilkan jawaban.

3.2 *Test results*

Pengujian sistem dilakukan dengan mengambil sinyal suara secara *real-time* menggunakan mikrofon. Dalam percobaan ini, sistem akan menghasilkan pertanyaan yang diucapkan pada robot. Penulis kemudian mengamati dan membandingkan *output* sistem dengan yang sebenarnya dengan pertanyaan yang diberikan oleh penutur yang digunakan untuk pengujian adalah "*where is the immigration office?*", "*where is the mayor's office?*", "*where is the airport?*", "*where is the nearest gas station?*", dan "*where is the nearest hospital?*". Untuk meningkatkan validitas hasil, sistem diuji dengan lima penutur secara acak, yang terdiri dari tiga pria kewarganegaraan Indonesia dan dua pria kewarganegaraan Prancis. Setiap penutur

mengucapkan lima pertanyaan sebanyak lima kali dengan total percobaan yang dilakukan adalah 125 kali.

Dapat diketahui bahwa nilai WER = 0.187 dari pengujian tersebut. Dalam pengujian tersebut, total substitusi kata adalah diperoleh (S) sebanyak 86, total penghapusan kata (D) sebanyak 5, dan penambahan kata (I) sebanyak 22. Prediksi kata pada pertanyaan yang mengandung kesalahan dijelaskan pada **Table 1**. Pada **Table 1**, pria kewarganegaraan Prancis pertama yang melakukan percobaan ditunjukkan dengan simbol FR1 dan pria kewarganegaraan Prancis kedua ditunjukkan dengan simbol FR2, begitu juga dengan pria kewarganegaraan Indonesia pertama ditunjukkan dengan simbol ID1, pria kewarganegaraan Indonesia kedua ditunjukkan dengan simbol ID2, dan pria kewarganegaraan Indonesia ketiga ditunjukkan dengan simbol ID3. Dari statistik yang diperoleh, "*where is the nearest gas station?*" adalah pertanyaan dengan tingkat kesalahan tertinggi. Terdapat empat kesalahan prediksi kata pada pertanyaan ini. Kata pada pertanyaan "*where is the nearest gas station?*" sering diprediksi sebagai "*guess*", "*news*", "*when*", atau "*new has*". Selain "*where is the nearest gas station?*", kata pada pertanyaan dengan tingkat kesalahan yang tinggi adalah "*where is the mayor's office?*". ASR sering kali mendeteksi kata pada "*where is the mayor's office?*" ini sebagai kata "*manager's*", "*main your*", dan "*his meal movies"*. Selain itu, ada pertanyaan "*where is the airport?*" yang diucapkan oleh FR1, yang diprediksi sebagai "*where is of half pot?*". Namun demikian, hal ini hanya terjadi satu kali dalam seluruh pengujian.

Table 1. List of errors in ASR prediction results during testing.

Person	Spoken question	Recognized question
ID1	where is the mayor's office?	where is the manager's office?
ID2	where is the mayor's office?	where is the main your office?
ID2	where is the nearest gas station?	where is the nearest guess the shin?
ID3	where is the nearest gas station?	where is the news gas station?
FR1	where is the nearest gas station?	when is the nearest gas station?
FR1	where is the airport?	where is of half pot?
FR1	where is the nearest hospital?	where is the newest was beaten?
FR2	where is the mayor's office?	where his meal movies?

Table 2. Samples of spoken question and the response time during testing.

No.	Person	Spoken question	Response Time (ms)		
			Fastest	Slowest	Average
1	ID1	where is the immigration office?	430 ms	720 ms	550 ms
2	ID1	where is the mayor's office?	510 ms	1640 ms	794 ms
3	ID1	where is the airport?	480 ms	650 ms	590 ms
4	ID1	where is the nearest gas station?	510 ms	760 ms	616 ms
5	ID1	where is the nearest hospital?	490 ms	650 ms	556 ms
6	ID2	where is the immigration office?	550 ms	700 ms	634 ms
7	ID2	where is the mayor's office?	490 ms	770 ms	584 ms
8	ID2	where is the airport?	440 ms	510 ms	484 ms
9	ID2	where is the nearest gas station?	530 ms	630 ms	578 ms
10	ID2	where is the nearest hospital?	500 ms	580 ms	548 ms
11	ID3	where is the immigration office?	260 ms	350 ms	298 ms
12	ID3	where is the mayor's office?	250 ms	310 ms	278 ms
13	ID3	where is the airport?	230 ms	310 ms	268 ms
14	ID3	where is the nearest gas station?	270 ms	390 ms	318 ms
15	ID3	where is the nearest hospital?	280 ms	350 ms	312 ms
16	FR1	where is the immigration office?	330 ms	1060 ms	536 ms
17	FR1	where is the mayor's office?	320 ms	530 ms	394 ms
18	FR1	where is the airport?	340 ms	430 ms	370 ms
19	FR1	where is the nearest gas station?	290 ms	490 ms	356 ms
20	FR1	where is the nearest hospital?	350 ms	490 ms	418 ms
21	FR2	where is the immigration office?	300 ms	540 ms	368 ms
22	FR2	where is the mayor's office?	330 ms	470 ms	400 ms
23	FR2	where is the airport?	360 ms	610 ms	430 ms
24	FR2	where is the nearest gas station?	310 ms	600 ms	438 ms
25	FR2	where is the nearest hospital?	410 ms	530 ms	458 ms

Setelah pengujian 125 kali, rata-rata waktu respons sistem yang dibutuhkan yaitu 464,04 *milliseconds*. **Table 2** menampilkan waktu respons sistem saat pengujian yang diperoleh dari 125 percobaan. Dalam pengujian ini, sistem menunjukkan waktu respons tercepat ketika penutur ID3 berbicara. Waktu respons terlama saat pengujian pada penutur ID3 adalah 390 *milliseconds*. Sedangkan, sistem menunjukkan waktu respons terlama saat pengujian ketika penutur ID1 berbicara. Percobaan dengan pertanyaan "*where is the immigration office?*" memiliki waktu respons yang cukup lama pada kelima penutur dilihat dari **Table 2**. Waktu respons sistem saat pengujian dipengaruhi oleh kecepatan dan keakuratan penutur saat berbicara.

4 Conclusion and future work

Penelitian ini menyimpulkan bahwa sistem QA menggunakan perintah suara pada robot humanoid telah berhasil diimplementasikan. Kesimpulan ini dibuktikan dengan beberapa pengujian yang dilakukan, baik pada ASR maupun sistem QA. ASR yang diusulkan pada penelitian ini memiliki performa WER sebesar 0.187. Nilai WER ini menunjukkan bahwa masih terdapat potensi kesalahan pengenalan kata pada sistem yang diusulkan. Penyebab kesalahan ini perlu diselidiki lebih lanjut, apakah berasal dari pengucapan yang salah oleh penutur *non-native* atau karena kekurangan pada ASR. Di masa mendatang, pengujian dengan

menggunakan penutur asli harus dilakukan untuk menyimpulkan hal ini. Dalam hal waktu respons, sistem memiliki waktu respons yang cukup lama dengan rata-rata sebesar 464,04 *milliseconds*. Peningkatan pada performa sistem QA perlu ditingkatkan agar memberikan pengalaman pengguna yang cepat dan responsif. Selain itu, di masa depan, sistem QA yang diusulkan pada robot *humanoid* dapat digunakan untuk mengembangkan robot layanan *humanoid* yang dapat berinteraksi langsung dengan manusia melalui percakapan.

Acknowledgments.

Dengan mengucapkan puji syukur atas nikmat yang diberikan Tuhan Yang Maha Esa. Penulis dapat menyelesaikan Tugas Akhir ini dengan lancar, di mana Tugas Akhir ini merupakan salah satu syarat mahasiswa untuk memperoleh gelar sarjana pada Program Studi Teknik Robotika. Ucapan terima kasih disampaikan kepada pembimbing saya, Bapak Eko Rudiawan Jamzuri, S.ST.,M.Sc, atas bimbingan dan arahan selama penelitian berlangsung. Saya juga ingin berterima kasih kepada teman dan penutur yang sudah bersedia meluangkan waktunya untuk membantu saya dalam penelitian ini.

References

[1] A. Ghosh, D. A. P. Soto, S. M. Veres, and A. Rossiter, “Human robot interaction for future remote manipulations in industry 4.0,” in *IFAC-PapersOnLine*, Elsevier B.V., 2020, pp. 10223–10228. doi: 10.1016/j.ifacol.2020.12.2752.

[2] L. Roda-Sanchez, T. Olivares, C. Garrido-Hidalgo, J. L. De La Vara, and A. Fernandez-Caballero, “Human-robot interaction in Industry 4.0 based on an Internet of Things real-time gesture control system,” *Integr Comput Aided Eng*, vol. 28, no. 2, pp. 159–175, 2021, doi: 10.3233/ICA-200637.

[3] A. Angleraud, A. Mehman Sefat, M. Netzev, and R. Pieters, “Coordinating Shared Tasks in Human-Robot Collaboration by Commands,” *Front Robot AI*, vol. 8, Oct. 2021, doi: 10.3389/frobt.2021.734548.

[4] A. Badr and A. Abdul-Hassan, “A Review on Voice-based Interface for Human-Robot Interaction,” *Iraqi Journal for Electrical and Electronic Engineering*, vol. 16, no. 2, pp. 1–12, Dec. 2020, doi: 10.37917/ijeee.16.2.10.

[5] M. Ünlü, E. Arisoy, and M. Saraçlar, *QUESTION ANSWERING FOR SPOKEN LECTURE PROCESSING*. 2018.

[6] D. Obeid, H. Ramambason, C. Pehlevan, and J. A. Paulson, “Structured and Deep Similarity Matching via Structured and Deep Hebbian Networks,” 2019.

[7] S.-W. ; Fan-Jiang, T.-H. Lo, and B. Chen, *SPOKEN DOCUMENT RETRIEVAL LEVERAGING BERT-BASED MODELING AND QUERY REFORMULATION*. IEEE, 2020.

[8] C. You, N. Chen, and Y. Zou, “Knowledge Distillation for Improved Accuracy in Spoken Question Answering,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 7793–7797. doi: 10.1109/ICASSP39728.2021.9414999.

[9] S. Abubakar *et al.*, *ARNA, a Service robot for Nursing Assistance: System Overview and User Acceptability*. 2020.

[10] H. S. Ahn *et al.*, "Hospital Receptionist Robot v2: Design for Enhancing Verbal Interaction with Social Skills," *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, doi: 10.1109/RO-MAN46459.2019.8956300.

[11] J. Ruby, S. Daenke, Y. Yuan, and W. Harry, "Automatic Speech Recognition and Machine Learning for Robotic Arm in Surgery Advanced Neurological MRI techniques View project 'Deep Robotics and Human Brain Simulation' View project," 2020. [Online]. Available: https://www.researchgate.net/publication/338375519

[12] B. S. Pranathi, A. Nair, C. S. Anushree, and T. S. Chandar, "SAHAYANTRA-A PATIENT ASSISTANCE ROBOT," 2020.

[13] T. Shimmura, R. Ichikari, T. Okuma, H. Ito, K. Okada, and T. Nonaka, "Service robot introduction to a restaurant enhances both labor productivity and service quality," in *Procedia CIRP*, Elsevier B.V., 2020, pp. 589–594. doi: 10.1016/j.procir.2020.05.103.

[14] N. T. Vo, P. V. Dang, T. N. Ngo, H. N. Le, and L. H. T. Do, *Restaurant Serving Robot with Double Line Sensors Following Approach*. IEEE International Conference on Mechatronics and Automation : IEEE ICMA, 2019.

[15] K. Berezina, O. Ciftci, and C. Cobanoglu, "Robots, artifiial intelligence, and service automation in restaurants," in *Robots, Artificial Intelligence and Service Automation in Travel, Tourism and Hospitality*, Emerald Group Publishing Ltd., 2019, pp. 185–219. doi: 10.1108/978-1-78756-687-320191010.

[16] M. Elazzazi, L. Jawad, M. Hilfi, and A. Pandya, "A Natural Language Interface for an Autonomous Camera Control System on the Da Vinci Surgical Robot," *Robotics*, vol. 11, no. 2, Apr. 2022, doi: 10.3390/robotics11020040.

[17] IBM, "What Is Speech Recognition? | IBM," www.ibm.com. Accessed: Sep. 27, 2023. [Online]. Available: https://www.ibm.com/topics/speech-recognition

[18] T. F. Pereira *et al.*, "A web-based Voice Interaction framework proposal for enhancing Information Systems user experience," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 235–244. doi: 10.1016/j.procs.2021.12.010.

[19] A. Cephei, "VOSK Offline Speech Recognition API," https://alphacephei.com/vosk/. Accessed: Sep. 27, 2023. [Online]. Available: https://github.com/alphacep/vosk-api

[20] I. OpenGenus, "RoBERTa: Robustly Optimized BERT pre-training Approach." Accessed: Dec. 14, 2023. [Online]. Available: https://iq.opengenus.org/roberta/

[21] H. Face, "roberta-base · Hugging Face." Accessed: Dec. 14, 2023. [Online]. Available: https://huggingface.co/roberta-base#Model%20Description

[22] D. Amos, "Python GUI Programming With Tkinter – Real Python," realpython.com. Accessed: Sep. 28, 2023. [Online]. Available: https://realpython.com/python-gui-tkinter/

[23] V. Sekhar, "Key Metrics for Evaluating Speech Recognition Software," Symbl.ai. Accessed: Sep. 27, 2023. [Online]. Available: https://symbl.ai/blog/key-metrics-and-data-for-evaluating-speech-recognition-software/https://symbl.ai/blog/key-metrics-and-data-for-evaluating-speech-recognition-software/

[24] T. Fukumori *et al.*, "Optical laser microphone for human-robot interaction: speech recognition in extremely noisy service environments," *Advanced Robotics*, vol. 36, no. 5–6, pp. 304–317, 2022, doi: 10.1080/01691864.2021.2023629.

[25] M. H. Korayem, S. Azargoshasb, A. H. Korayem, and S. Tabibian, "Design and Implementation of the Voice Command Recognition and the Sound Source Localization System for Human-Robot Interaction," *Robotica*, vol. 39, no. 10, pp. 1779–1790, Oct. 2021, doi: 10.1017/S0263574720001496.

[26] K. Zagalo, O. Verbytska, L. Cucu-Grosjean, and A. Bar-Hen, "Response Times Parametric Estimation of Real-Time Systems," Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.01720