Journal of Biomedical Informatics 144 (2023) 104446

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automating medical simulations

Sapir Gershov [a,*], Daniel Braunold [b], Robert Spektor [c], Alexander Ioscovich [d], Aeyal Raz [b], Shlomi Laufer [c]

[a] Technion Autonomous Systems Program, Technion - Israel Institute of Technology, Haifa, Israel
[b] Rambam Health Care Campus, Haifa, Israel
[c] Faculty of Industrial Engineering & Management, Technion - Israel Institute of Technology, Haifa, Israel
[d] Shaare Zedek Medical Center, Jerusalem, Israel

ARTICLE INFO

Keywords:
Medical Simulation
Checklist
Activity Recognition
Medical
NLP
Machine Learning
Speech

ABSTRACT

Objective: This study aims to explore speech as an alternative modality for human activity recognition (HAR) in medical settings. While current HAR technologies rely on video and sensory modalities, they are often unsuitable for the medical environment due to interference from medical personnel, privacy concerns, and environmental limitations. Therefore, we propose an end-to-end, fully automatic objective checklist validation framework that utilizes medical personnel's uttered speech to recognize and document the executed actions in a checklist format.
Methods: Our framework records, processes, and analyzes medical personnel's speech to extract valuable information about performed actions. This information is then used to fill the corresponding rubrics in the checklist automatically.
Results: Our approach to activity recognition outperformed the online expert examiner, achieving an F1 score of 0.869 on verbal tasks and an ICC score of 0.822 with an offline examiner. Furthermore, the framework successfully identified communication failures and medical errors made by physicians and nurses.
Conclusion: Implementing a speech-based framework in medical settings, such as the emergency room and operation room, holds promise for improving care delivery and enabling the development of automated assistive technologies in various medical domains. By leveraging speech as a modality for HAR, we can overcome the limitations of existing technologies and enhance workflow efficiency and patient safety.

1. Introduction

In the medical field, developing medical residents' clinical skills necessitates interaction with real-life patients. However, this imperative must be balanced with the ethical obligation to ensure patient safety[1]. Consequently, there is a growing interest in establishing performance-based assessments for medical practitioners that demand a demonstration of competence [2]. From that need, simulation-based assessment has emerged as a promising approach for evaluating hands-on skills, knowledge, clinical reasoning, communication skills, decision-making, and teamwork [2–5].

Qualitative and quantitative performance-based validation metrics have been developed to ensure that simulation-based assessment is a reliable examination method [6]. Among these metrics, task-specific checklists have gained popularity due to their simplicity. These checklists provide step-by-step guidance and criteria for assessing observable behaviors, offering a relatively intuitive evaluation method [7]. A comprehensive simulation task-specific checklist is designed by expert evaluators who outline the crucial actions candidates should perform, based on presented vitals and symptoms, to manage the scenario effectively [8,9]. However, this form of assessment has a significant drawback - the simulation examiner must complete the checklist while the simulation is being executed. This increases the possibility of human error arising from limited visual and auditory perception [10]. Moreover, this approach is expensive and time-consuming, as it requires the involvement of professional physicians [4,9,11,12]. Finally, it does not provide medical practitioners with the opportunity for independent training.

Ideally, the checklist completion process should be automated using an objective framework capable of recognizing, monitoring, and documenting medical residents' performance. Unlike human evaluators, machines can record and process abundant amount of visual and auditory data, enabling them to minimize the costs associated with performance assessments and allowing more residents to train in complex

* Corresponding author at: Technion Autonomous Systems Program, Technion - Israel Institute of Technology, 32000 Haifa, Israel.
E-mail address: sapirgershov@campus.technion.ac.il (S. Gershov).

https://doi.org/10.1016/j.jbi.2023.104446
Received 19 December 2022; Received in revised form 8 July 2023; Accepted 16 July 2023
Available online 17 July 2023
1532-0464/© 2023 Elsevier Inc. All rights reserved.

scenarios independently.

Adequate information sharing through verbal communication among medical personnel is crucial, particularly during complex resuscitations and medical emergencies [13]. Furthermore, simultaneous collaboration and reporting progress to the supervising physician are imperative to manage medical emergencies successfully. Therefore, we hypothesized that medical personnel communication embodies information about the executed actions.

Our prior work [14] underscored the importance of speech as a robust cue for activity recognition. We presented preliminary results of a speech-based framework capable of identifying medical activities, which provided valuable insights for designing a comprehensive speech-based activity recognition system for the medical domain.

2. Related work

Human Activity Recognition (HAR) involves identifying distinct actions or movements based on data captured by sensors. Much of the existing research emphasizes on the recognition of visually observable actions [15–24]. However, it's crucial to acknowledge that certain activities, such as verbal interactions, do not necessarily depend on visible movements or objects for their identification. Therefore, audio data can be a potent indicator for recognizing a broad spectrum of activities [24–27].

Within the medical field, the integration of HAR technology presents unique hurdles due to the fast-moving, multitasking nature of clinical environments [25,28–30]. Prior studies have investigated modalities like RFID, wearable sensors, and video for recognizing clinical activities. RFID and wearable sensors present benefits in terms of cost and size, but their effectiveness is hampered by environmental factors that interfere with signal reception, the limited number of objects that can be tagged, and the material properties of medical instruments. The use of video, while helpful, presents limitations such as privacy issues for patients, potential obstruction of camera views by medical personnel and equipment in congested areas.

Previous research has explored the potential of speech-based HAR in the intricate contexts of medical emergencies [25,26,31,32]. These studies focused on recognizing activities' types and performance stages through verbal communication in real-life hospital environments, providing insights into the challenges associated with designing speech-based activity recognition frameworks.

These papers solely focus on activities that are commonly performed during trauma resuscitation. Second, these works only utilized keyword-based methods that rely on a predefined keywords list.

In contrast to these papers, our work implements an end-to-end speech-based HAR for evaluating anesthesia residents clinical skills in a simulation environment. Medical simulations provide a controlled training and assessment environment that emulates real-life settings while minimizing disturbances and allowing examinees to speak directly and fluently without the stress of endangering patients. In addition, we chose a more direct approach to record the medical personnel's speech, thus reducing interference and background noise. Furthermore, we developed a sentence-level methodology for identifying activities from medical members spoken language. This approach is more suitable for analyzing human language and eliminates the need for a predefined keywords list.

Our work introduces several methodological innovations and contributions. Firstly, we investigate the feasibility of speech-based activity recognition in the dynamic clinical domain. Compared to standard modalities like computer vision and RFID, speech is better suited for preserving patient privacy and addressing the sparsity of medical emergencies. Additionally, speech can provide robust cues for activities that are not solely based on actions or objects. This contributes to the literature by (1) characterizing the benefits and challenges of using speech as a modality for action recognition in the medical domain and (2) presenting a comprehensive pipeline for action recognition from raw audio recordings. The insights gained from this research will guide the development of an automatic activity recognition system suitable for real-life hospital settings.

Secondly, we propose a novel algorithm for sentence matching. Traditional methods relied on word-level approaches, such as lexical and semantic similarity, to estimate the relationship "strength" between text corpora through a numerical description obtained according to the word level comparison [33–36]. Nonetheless, these methods are less effective when significant variations exist between sentences, as often found in free speech. To address this challenge, we explore the potential of pre-trained language model sentence embeddings. By leveraging pre-trained language models like "BERT" [37], we generate fixed-length, multi-dimensional embeddings that capture the meaning of the represented sentence. We specifically employ the cross-encoder structure of language models and incorporate a classification head to enable sentence-level analysis and extract more informative information.

The utilization of sentence-based analysis opens new possibilities for examining medical communication. Ineffective communication among medical professionals is a leading cause of medical errors and patient harm, particularly in acute care units where managing communication becomes more challenging [38–40]. In an attempt to overcome this challenge, physicians execute the "Readback" protocol – a procedure whereby the receiving station repeats a received message to ensure a safe closed-loop communication between medical personnel [41]. Our work introduces an innovative approach for detecting inappropriate instances of "Readbacks" using the proposed sentence similarity calculation methodology. This approach successfully identifies cases that may have been overlooked by human examiners and holds potential for integration into hospital systems in the future.

3. Materials

3.1. Participants

The Israel Society of Anesthesiologists organizes a preparation day to allow anesthesiology residents to practice in high-fidelity simulations and receive feedback from experienced anesthesiologists. This study focuses on analyzing the performance of participants during these preparation days.

Forty-seven senior anaesthesiology residents participated in the study, comprising 34 (72.3%) males and 13 (27.7%) females. The residents were recruited from 24 different hospitals, representing a diverse range of medical simulation experience. This diversity allows us to train and evaluate the framework's robustness using a comprehensive dataset.

Each resident was recorded while managing up to two different simulation scenarios. The dataset used for training and evaluating our framework can be found in the supplementary file (Table S1).

Two research team members assumed the roles of a nurse and a medical intern and assisted the examinees, based on their orders, to enhance the realism of the scenarios. Additionally, an experienced anesthesiologist evaluated the residents' performance using a scenario checklist and provided feedback after the examination.

Before participating in the study, all participants signed an informed consent form and completed a demographic questionnaire, including questions about their previous experience with medical simulations. The Rambam Medical Center IRB committee approved the study.

3.2. Medical simulations

The framework's evaluation involved assessing senior anesthesia residents' skills using five simulation scenarios. An experienced anesthesiologist and a medical simulation expert collaboratively developed these scenarios. Additionally, two scenarios were derived from previous Israeli Anesthesiology Board certification simulation exams. The simulation scenarios encompassed the following cases: (1) the management

of a patient with a severe anaphylactic (allergic) reaction, (2) a post-surgery patient experiencing severe bradycardia (slow heart rate), (3) a post-operative patient suffering from opiate overdose, (4) a post-operative patient suffering from opioid overdose, and (5) a post-operative patient experiencing severe hypoglycemia (low blood sugar level).

As done in similar medical simulation studies [42–45], a detailed checklist was created for each scenario to ensure a comprehensive assessment. Each checklist comprised an average of 35 tasks, which were evaluated based on the quality of their execution following established medical guidelines. The score for each checklist task ranged from 0 to 2, with the following scale: 0 for tasks not observed, 1 for tasks requiring improvement, and 2 for tasks meeting expectations.

3.3. Data acquisition system

For the simulation's patient role, we utilized two full-body manikins designed explicitly for Advanced Cardiovascular Life Support (ACLS) training: MegaCode Kelly and SimMan 3G, manufactured by 'Laerdal'.

To capture the simulations comprehensively, we recorded video and audio using StreamPix digital video recording software (NorPix Inc.). Subsequently, an independent human observer reviewed the recorded video data to validate the checklist assessments made by an experienced anesthesiologist.

For audio recordings, all simulation participants (resident, nurse, and examiner) wore wireless lavalier microphone transmitters connected to a digital mixer. Each audio channel was saved separately to minimize recording disturbances and ensure clear audio capture.

The simulations were conducted within the hospital's post-anesthesia care unit (PACU) and incorporated real-life medical equipment and utilities, creating an authentic training environment. (Fig. 1.).

4. Methods

4.1. End-to-end automatic checklist completion

After the data collection, the checklist tasks are completed using the proposed pipeline, which consists of several stages. Firstly, we performed preprocessing on the acquired data to remove any disturbances in the recording and enhance the efficiency of subsequent steps. Next, we employed Google's speech-to-text API to transcribe the processed recordings automatically. Subsequently, we identified keywords within each transcribed sentence. Finally, we implemented a matching process between the checklist tasks, the specified keywords, and the sentences from the transcription corpus.

The algorithm produces a filled-out checklist where the tasks identified by the framework are associated with a matching sentence from the transcription and its timestamp. For a more comprehensive understanding of the process please refer to Fig. 2, which provides a detailed description.

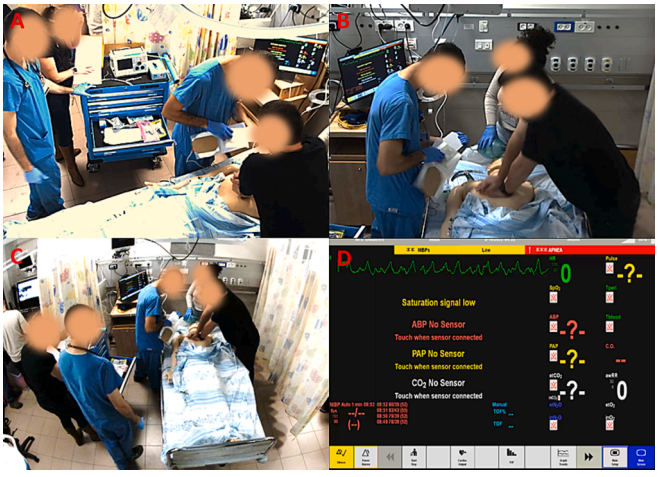


Fig. 1. Data acquisition system. (A) Nurse working area; (B) Physician working area; (C) Overview of the simulation area; (D) Patient monitor. The blurred faces in this figure are human-generated.

4.2. Data preprocessing

Raw audio recordings containing severe audio-visual disturbances are unsuitable for accurate automatic transcription using Google's speech-to-text API. Therefore, this stage also filters out simulation recordings that are incomplete or contain severe artifacts, thereby removing them from the dataset. The preprocessing stage consists of three steps:

1. **Speech Source Separation.** Due to the limited space in the simulation area, the simulation participants (resident, nurse, and intern) are in close proximity to each other, resulting in overlapping audio signals. Additionally, background noises from patient monitors and defibrillators are present in the recordings. To address this, we applied a source separation algorithm (commonly referred to as the 'cocktail party problem' [46]) on each raw audio recording, separating the mixed signal into a set of source signals without additional information about the sources or the mixing process [47].

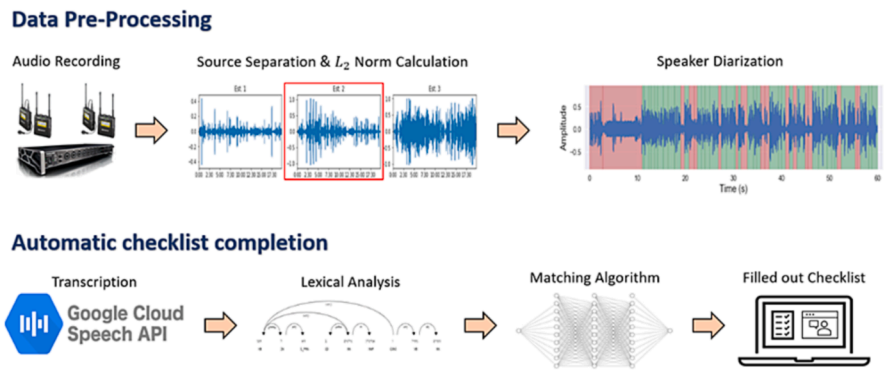
   In the past few years, several open-source projects have implemented speech source separation using deep learning methods [48–50]. For this study, we implemented the Conv-TasNet [51] network provided by Asteroid [48], trained initially on English speech. We performed fine-tuning based on the work of Anidjar et al. [52] to adapt it to Hebrew speech and the specific recording settings of the simulations. For each team member the network outputs three audio channels' which represent the main speaker, the secondary speaker, and background noise. Thus, a total number of nine audio files are generated.
2. **Primary File Identification.** Instead of transcribing all nine audio files, which is inefficient, we simplified the process by treating the audio files as a vector and calculated the $L2$ norm for each recording. Afterward, we chose the primary file based on the highest norm values, indicating speech signals' presence. The selection process is repeated for each team member.

   It Is important to note that determining the primary file for the nurse audio channel is more challenging. It is likely due to the significantly less frequent speech intervals compared to the physician examinee.
3. **Speaker Diarization.** Speaker diarization involves determining "who spoke when" in an audio recording and assigning labels to audio segments corresponding to speaker identities [53]. Machine learning techniques have been successfully applied to speaker diarization, contributing to state-of-the-art performances in automatic speech recognition frameworks [53–55]. In this work, our goal was to provide a complete transcription of the simulation recording with timestamps for each sentence. We utilized the 'pyAudioAnalysi' library [56], incorporating an SVM model for semi-supervised audio segmentation. The model takes an uninterrupted speech recording as input and outputs endpoints corresponding to "silence" areas, aiding speech segmentation accuracy. In addition, we performed fine-tuning of the thresholds to improve the division accuracy.

4.3. Morphological and syntactic parsing

When dealing with Morphologically Rich Languages (MRLs) like Hebrew, traditional syntactical analysis faces a significant challenge due to the high degree of morphological ambiguity. However,

the process of lemmatization which converts input tokens into their constituent morphemes [57], can effectively address this challenge [58].



Task	Timestamp	Evidence in Text
Listening to the lungs and heart	00:03:00.520	"I am listening."
Adrenaline 0.2 - 0.3 mg	00:04:28.040	"Give me Adrenaline. OK, thank you. Adrenaline is inside."
Ventolin Inhalation - Half a cc of Ventolin	00:03:54.540	"We can do Ventolin inhalation."

Fig. 2. Automatic checklist process. End-to-end description of the checklist completion pipeline and the generated output.

In their paper [59], Tsarfaty et al. introduced a parser that leverages global context to accurately decompose raw Hebrew tokens into their respective morphemes. We utilized their parser in our study to reduce the variance in the transcription database, thereby improving the results of the matching algorithm.

Please refer to the supplementary file (Figure S1) for a visual representation of MRL parsing.

4.4. Matching algorithm

In our previous work [14], the matching algorithm utilized a single heuristic to assess the similarity between a transcribed sentence and a checklist task. In this study, we proposed a more advanced algorithm to provide the framework with more accurate information. The algorithm applied two granulation levels to analyze the simulation participants speech and found the most suitable match for a task in the checklist. This process consists of two components:

1. **Word Importance Similarity – similarity on a word level**

   To guide the examiner in identifying the different assignments, each task in the checklist includes a brief description of the expected treatment. After applying lemmatization to each task description, we selected specific words that best represent the task and generated a "bag-of-words" representation. These keywords typically include medical terms, medications, procedures, instruments, and combinations of objects and verbs.

   To measure the similarity between a task description and a sentence in the transcription, we used the term frequency-inverse document frequency (TF-IDF) [60] approach with a threshold. TF-IDF is a numerical statistic that reflects a word importance to the sentence central theme. By calculating the TF-IDF score for each word in the task description, we can then choose the word with the highest value as the TF-IDF threshold of the current task. In this work, we choose two words for each task where the lower value is the threshold. We then merged all team members transcriptions, three in total, into a single corpus based on the timestamps of each sentence and calculated the TF-IDF score for each word in a sentence. Finally, a sentence will be considered a suitable match if it contains at least one of the two words of the task and at least one of them has a TF-IDF value that is equal or larger than the threshold.

2. **Sentence Embedding Similarity – similarity on a sentence level**

   Semantic Textual Similarity (STS), the task of determining the semantic equivalences between two textual contents, is considered a daunting research challenge. This problem becomes particularly significant for under-resourced languages such as Hebrew [36,61]. Historically, STS methodologies hinged on knowledge-based, corpus-based, and deep neural network approaches [35]. Nonetheless, the recent surge in machine learning advancements brought forth language models – transformer-based models that utilize sentence embedding methods and similarity measurements like cosine similarity or Euclidean distance. These models surpassed their predecessors and achieved unprecedented results in multiple NLP domains [37,62–64].

   In our study, we harnessed the capabilities of AlephBERT [58], a language model trained specifically on Hebrew corpora. However, directly applying language models to niche domains, such as the medical field, may not always result in optimal outcomes due to the domain-specific language. To tackle this, we applied domain adaptation, which necessitates the fine-tuning of AlephBERT using a masked language model task. During this process, we obscured tokens (usually words) in a sentence, specifically technical terms and jargon not previously encountered during the model training, and then guided the model to replace the mask with a suitable token. This helps the model to gain a contextual grasp of the entire sentence. Post fine-tuning, we computed the similarity between sentence vectors using cosine similarity. This involves channeling the embeddings from the fine-tuned model output layer through a pooling operation, producing fixed-sized vectors for each input sentence. We then deployed the mean squared-error loss as our training objective function. (Fig. 3.).

4.5. Checklists comparison

An independent assessor conducted a secondary analysis and filled out the checklist again. Similar to the online examiner, the offline evaluator used a scoring range from 0 to 2 for each task on the checklist.

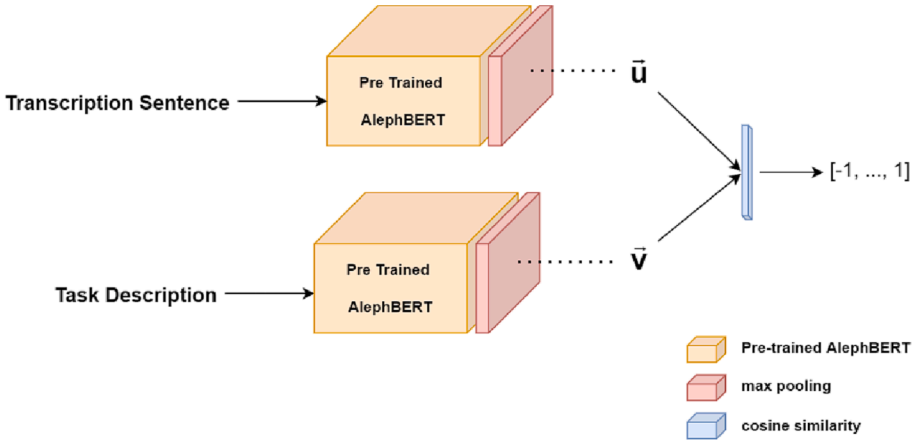


Fig. 3. Model architecture to compute similarity scores. This architecture is also used with the regression objective function for fine-tuning our dataset.

Unlike the human evaluators who assess the quality of execution, our algorithm simply verifies the task completion, thus attributing a binary score. This necessitated a similar binary scoring approach for the human-completed checklists.

To compare the outputs of our framework with the checklists completed by humans (both online and offline), we adopted the F-score [65] as our evaluative metric. The F-score, also referred to as the F_1-score, is frequently used to assess binary classification models, computing the harmonic mean of precision and recall.

In addition, we gauged the concurrence between human evaluators and our algorithm using the Intraclass Correlation Coefficient (ICC) score [66]. The ICC, a universally accepted method for gauging rating reliability, ranges from 0 to 1, with a higher ICC score nearing 1 signifying greater similarity within the same group, and a lower score indicating lesser similarity. For our study, we computed the ICC score using one-way random effects, absolute agreement, and multiple raters, represented as ICC(1, k).

4.6. Evaluation of the matching algorithm

To assess the performance of our framework pipeline, we will evaluate the matching algorithm in two stages. First, we will examine the hypothesis that the algorithm can effectively document the executed actions. Second, we will compare the filled-out checklists from all three evaluators: the online evaluator, the offline evaluator, and the framework.

Our underlying assumption is that the offline evaluator, who has no time limitations and can review the video multiple times from different angles, will produce the most accurate checklist.

To validate the framework capability of accurately documenting the executed actions, we will compare its performance with other algorithms designed for similar tasks. This comparison will help us assess the effectiveness and accuracy of our framework in capturing the relevant actions during the simulation.

In addition, we performed a more detailed analysis by dividing the different checklist actions into five major categories:

- Diagnosis– Asks for information, recognizes condition, asks for patient chart and labs, etc.
- Physical exam – Physical examination of the patient, i.e., listening to the heart and lungs.
- Medication management – Request specific medication and dosage.
- Medical equipment – Requests for a defibrillator, oxygen mask, ECG, etc.
- Procedure management – Follows ACLS protocol, ROSC protocol, call for help, etc.

4.7. Recognition of communication failures

In the available dataset, events of "Readback" failures are marked and divided into two groups: administering a wrong medication and administering an improper dosage. For this analysis, we assumed that the nurse would perform a "Readback" as soon as possible and with minimal variation from the physician's original sentence. For each labeled "Readback" in the appropriate and inappropriate dataset, we calculated the absolute value of the cosine similarity between the sentences and converted it into a percentage. We speculated that the differences in similarity scores are significant and can be used as a classification criterion. To validate our hypothesis, we applied Welch t-Test [67]. This statistical test helps determine whether there is a significant difference between the means of two groups, even when the variances are unequal. Based on the results of this test, we developed an algorithm to detect inconsistencies in the "Readback" dialogues.

5. Results

5.1. Evaluation of the matching algorithm

We compared the impact of different matching algorithms on the framework performance, as shown in Table 1. As it shows, our proposed algorithm achieved the highest F_1 Score.

In addition, after differentiating the checklist actions into 5 major classes, we received a more detailed analysis of the matching algorithms' impact (Table 2):

5.2. Evaluation of audio channels impact on accuracy

Another aspect of the framework that we evaluated was the impact of each simulation participant on the framework's accuracy. We show that each participant in the simulation contributes essential information via

Table 1
Ablation study of the algorithm's performances over the collected data. This evaluation considered the number of tasks recognized by the algorithm and the appropriate F_1 score.

Algorithm	Total Tasks	Tasks Executed	Framework Identified	F_1 Score
Naïve approach	2513	2079	1833	0.794
TF-IDF similarity			1871	0.803
Sentence embedding similarity			1874	0.810
TF-IDF & Sentence embedding similarity			**1903**	**0.834**

Table 2
Ablation study of the algorithm's performances over the collected data. This evaluation granulates the recognized tasks based on category. The reported values are the F_1 scores. N - the number of executed tasks; n – the number of identifications made by the human observer.

Algorithm	Diagnosis $N = 410;\ n = 330$	Physical exam $N = 375;\ n = 270$	Medication management $N = 761;\ n = 729$	Medical equipment $N = 356;\ n = 344$	Procedure management $N = 177;\ n = 126$
Naïve approach	0.761	0.561	0.910	0.929	0.542
TF-IDF similarity	0.753	0.592	0.914	0.958	0.577
Sentence embedding similarity	0.724	0.605	0.908	0.944	0.652
TF-IDF & Sentence embedding similarity	**0.804**	**0.717**	**0.947**	**0.953**	**0.706**

speech, which is necessary for the human evaluator as much as it is for our framework. This implies that while monitoring the examinee's performance, the examiner must also listen carefully to other participants. This is clearly indicated by improved framework performance with additional audio channels (see Table 3).

The following table presents the impact of each participant recording on the framework accuracy measured by F_1 Score:

5.3. Evaluation of the framework performance in comparison to the human evaluator

In evaluating the framework as a replacement for the online human examiner, we assessed the level of agreement among three evaluators: the online human evaluator, the offline human evaluator, and our algorithm.

On the one hand, a framework that utilizes all audio channels has a greater agreement with the offline evaluator. While on the other, a framework that only relies on the physician's audio channel has a greater agreement with the online evaluator. This result is indicated in the following table:

5.4. Assessment of communication failures

In the evaluation of sentence similarity, the first step aimed to establish a baseline value for a random pair of sentences. This involved calculating the absolute value of the cosine similarity score for each combination of sentence pairs in the complete dataset. The calculation was done without considering the context or time difference between the sentences. The average score obtained from this calculation was $17.4\% \pm 15.8$.

Based on the available "Readback" dataset, the following results are generated:

As we mentioned earlier, we applied Welch's t-Test on the "Readback" dataset to validate our hypothesis, which resulted in a test statistic of 18.508 and the corresponding p-value is $5.415e^{-10}$.

6. Discussion

Most literature works in the field of HAR have reported promising results by focusing on visual and sensory data. However, while there are some activities that are not visually detectable, they can be verbally reported. Furthermore, these modalities are unsuitable for the multitasking, fast-paced, and concurrent processes of clinical activities. Moreover, visual-based activity recognition in the medical domain is considered problematic because of patient privacy concerns and frequent obstruction of camera views.

Table 3
The effect of each member's speech recording on the framework accuracy measured by F_1 score.

Single Audio Channel			**Multi Audio Channels**			
Physician	Nurse	Intern	Physician & Nurse	Physician & Intern	Nurse & Intern	Physician & Nurse & Intern
0.683	0.588	0.122	0.774	0.755	0.611	**0.834**

Table 4
All three evaluators' ICC scores based on one-way random effects, absolute agreement, multiple raters - ICC (1, k).

			Evaluator		
		Audio channel	Online	Offline	Framework
Evaluator	Online	All Audio Channels	1	0.655	0.787
	Offline		0.655	1	0.822
	Framework		**0.787**	**0.822**	1
	Framework	Physician Audio Channel	**0.849**	**0.734**	1

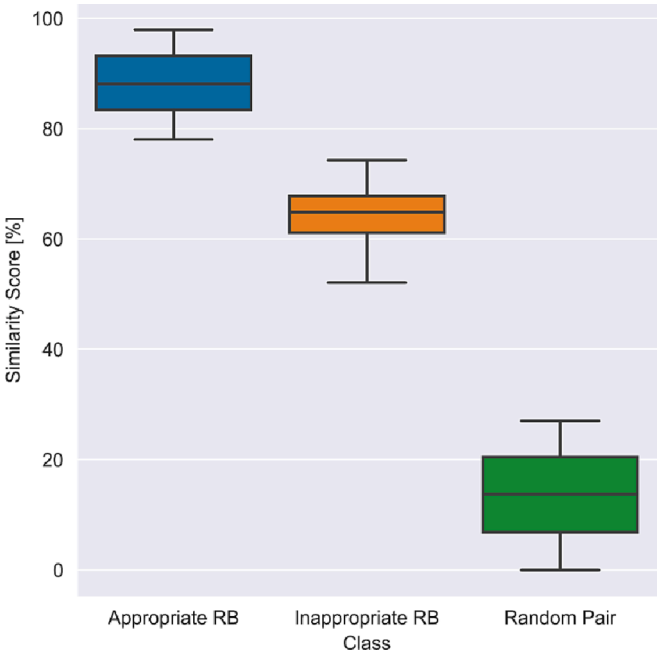


Fig. 4. Illustration of the communication failures statistical results. A Boxplot of the sentence's similarity scores from the labeled "Readback" dataset and random pairs from the available dataset. RB – Readback.

Previous works that explored the potential of speech-based HAR in the medical domain, have focused on recognizing activities' that are commonly performed during trauma resuscitation. These works utilized keyword-based methods that rely on a predefined keywords list and a manual transcription of the team verbal communication.

A comparison between these works and ours is located in the supplementary file (see Table S3).

In contrast to these papers, our work implements speech-based HAR for evaluating clinical skills in a simulation environment. More specifically, we introduce an end-to-end, fully automatic framework for objective checklist validation. By exploiting the analysis of medical professionals uttered speech, this framework can efficiently and accurately fill in relevant sections of a checklist, by harnessing the power of

sophisticated machine learning algorithms and techniques.

We initially evaluated various audio preprocessing parameters to identify optimal audio segments. Subsequently, we transcribed the simulation audio recordings using these parameters and employed a multi-level similarity-based matching algorithm. Finally, a performance assessment of the algorithm was conducted using different similarity metrics.

It is important to note that while our framework can process several speakers without damaging its performance, it may not fully capture the residents' nuanced actions. Nonetheless, we believe this end-to-end, speech-based, fully automatic framework for assessing medical residents is the first of its kind, representing a significant advancement in the field.

Another crucial aspect of the framework is its ability to identify medical errors associated with communication failures among physicians and nurses, a vital factor for ensuring patient safety. Our observations indicate that human examiners occasionally overlooked such errors. To address this issue, our framework incorporates the evaluation of "Readback" procedures, harnessing the potential of transformer-based language model such as AlephBERT [58].

The successful creation and utilization of this framework carries substantial implications for the healthcare industry. This kind of audio-driven framework could find practical application in real-world medical environments like emergency rooms and operating rooms. By offering automated checklist validation based on audio cues, this framework can streamline and boost the monitoring and evaluation process of medical procedures. Furthermore, this framework holds promise for the inception of automated assistive technologies across a broad range of medical sectors. By capitalizing on advanced speech analysis techniques, analogous frameworks could be developed to aid medical professionals in various tasks such as documentation, decision-making, and quality assurance. Such prospective technological developments could dramatically enhance efficiency, precision, and patient safety.

6.1. Evaluation of the matching algorithm

The results in Tables 1 & 2 show that the TF-IDF similarity and the sentence embedding similarity algorithms outperformed our original algorithm. Furthermore, when both methods were combined in the checklist-filling process, there was a significant improvement in the framework's results across all categories. This demonstrates that the new framework is more suitable for autonomously documenting the actions performed during clinical simulations.

The benefits of using AlephBERT embeddings as a method for sentence comparison will be discussed further in a dedicated section, providing more detailed insights into its effectiveness.

Table 4 indicates that the TF-IDF similarity algorithm achieved better results in specific categories than the sentence embedding similarity algorithm. This is primarily due to particular tasks that can be classified as "executed" based on a single word. For instance, in the condition diagnosis task, where the examinee is expected to explicitly state the patient's condition (e.g., Anaphylaxis, VF), the examiner may consider the task accomplished even if the participant did not specifically use the checklist-required words. The online evaluator might consider phrases from the same semantic field (e.g., "allergies" and "sensitivities") as suitable alternatives to the word "anaphylaxis," which reflects the examinee's accurate decision-making.

While the sentence embedding similarity algorithm is effective in many cases, the TF-IDF similarity algorithm is more suitable for tasks where a single word carries significant meaning and indicates task completion. These findings highlight the importance of considering contextual information and semantic relationships when assessing the execution of specific tasks.

Overall, these findings contribute to understanding the strengths and limitations of different algorithms in the framework and provide valuable insights for further refining the framework's performance and enhancing its accuracy in documenting executed actions during clinical simulations.

6.2. Evaluation of the framework as a suitable replacement for the online human examiner

Our underlying assumption is that the offline evaluator is the most accurate due to the ability to re-examine simulations with no time limitations and from multiple perspectives is reasonable. Thus, considering the offline evaluator as the ground truth in this experiment provides a reference point for evaluating the accuracy of other evaluators, including the online evaluator and the framework.

Previous works have reported inferior ICC scores among online evaluators [68–70], which indicates that relying solely on the online evaluator may lead to inadequate documentation and evaluation of medical performance. In addition, these findings further support the need for a robust objective checklist validation framework.

As presented in Table 4, the algorithm developed in this study demonstrates higher reliability and lower error rates compared to the online evaluator. This suggests that the framework can serve as a suitable alternative to an experienced online evaluator. However, there is still room for improvement in terms of ICC scores, indicating the potential for further refinement and enhancement of the framework.

It is worth noting that providing the algorithm with only the examinee recordings resulted in improved ICC scores with the online evaluator. However, this came at the expense of reduced ICC scores with the offline evaluator and less accurate F_1 scores on the action recognition task. This finding suggests that the online evaluator primarily focuses on the examinee, leading to disagreements with the offline evaluator.

Since the goal is to emulate the performance of the offline evaluator, who is likely to be more accurate, the algorithm described in this paper is designed to process the recordings of all participants before making a final judgment. Thus provide a reliable and consistent assessment of medical performance.

6.3. Assessment of communication failures

Fig. 4 demonstrates the application of the AlephBERT model in pinpointing communication failures. The plot exhibits the similarity scores among sentences, facilitating the establishment of a classification threshold vital for differentiating between appropriate and inappropriate "Readback" dialogues. To confirm these findings, we performed Welch's *t*-test, yielding a p-value that endorses the framework's reliability.

6.4. BERT embeddings as a method for sentence comparison

Until recently, neural network models generated word embeddings that remained fixed regardless of the surrounding context. On the other hand, BERT was among the first models to introduce dynamic word representations influenced by the neighboring words. This contextual information allows BERT to capture broader linguistic nuances and improve the model's overall performance.

It is important to note that direct word-level similarity comparisons using BERT embeddings are unsuitable due to their contextual nature. The embedding of a word varies depending on the sentence it appears in. However, comparing sentence embeddings is still a valid approach. By employing a superficial similarity metric, we identified sentences with contextual relevance, which was instrumental in assessing the similarity between checklist task descriptions and the transcriptions of participants. Furthermore, the fine-tuned AlephBERT model demonstrated its ability to rank sentences based on similarity, even when dealing with subtle differences. This was evident in distinguishing between appropriate and inappropriate "Readback" dialogs, where minor variations in sentence wording played a crucial role.

As technology continues to advance, Large Language Models (LLMs)

have emerged as a groundbreaking innovation in the field of NLP. These models, such as GPT-4 [71], have the ability to generate human-like text, revolutionizing natural language processing tasks. And specifically in the medical domain [72]. However, these models require massive computational resources. Thus, in this study we used AlephBERT which provided meaningful embedding and may be executed locally using reasonable resources.

6.5. Limitations

One limitation of our methodology is the algorithm's inability to accurately identify tasks that are associated with procedure management and physical exams. These categories often involve actions that are predominantly performed without verbal communication, such as undressing the patient, taking vital measurements, or executing the ACLS protocol. Since our algorithm relies on analyzing speech, these non-verbal tasks are undetectable.

To address this limitation, one potential solution is to incorporate data fusion by combining verbal communication with data from the simulation manikin. For example, SimMan 3G can detect physical procedures like pulse measurements and chest compressions. By integrating such data, we can compensate for cases where speech is not expected or required.

Additionally, when performing physical exams on real-life patients, physicians usually verbally communicate with the patient, explaining the specific exam they will conduct (e.g., "I will now listen to your lungs"). Regrettably, the incorporation of a simulated patient may have induced a feeling of unrealism among some participants. This might have led them to deviate from standard practices, thereby adding an extra layer of complexity to the process of speech-based recognition.

Another significant challenge is the nuanced nature of residents' actions. While speech provides evidence of actions being executed, it offers limited information regarding their quality. The quality of an action can significantly impact whether a checkbox is satisfied or not. To overcome this challenge, incorporating additional modalities of data, such as video recordings and sensory data, can enrich the dataset and provide more accurate predictions that go beyond what can be inferred from a speech-based framework alone.

6.6. Future work

We plan to continue our research in the medical domain and explore the applicability of language models further. Improving the model's comprehension of "Readback" sentences is a valuable direction to pursue, as it can enhance the accuracy and effectiveness of the framework.

In addition, investigating sentiment analysis of both speech and text in the clinical setting is also a promising avenue for research. Sentiment analysis can provide insights into the emotions and attitudes expressed by medical personnel and patients, which can be valuable for decision-making and improving overall healthcare outcomes. Expanding the research in this area can contribute to developing tools and models that aid in understanding sentiment in clinical interactions.

Incorporating the results of our project as part of a machine-learning model to assess resident performance aligns with the long-term objective of leveraging technology for performance evaluation in medical education. Thus, by integrating our findings into a comprehensive machine-learning model, we can contribute to developing objective and data-driven approaches for assessing and enhancing the skills of medical residents.

Ethical considerations

It is crucial to clarify that we gathered valuable insights and evaluated the framework's performance using simulated scenarios without exposing actual patients to potential risks. In addition, we have ensured compliance with ethical standards and guidelines for human experimentation, including adherence to the Helsinki Declaration.

Any document of harmful actions has been removed from our database during the development of the proposed framework. This ensures that the framework operates within safe and acceptable boundaries.

The ability of our autonomous model to incorporate physician input without interfering with their work is a significant advantage. This allows the model to leverage the expertise and insights of physicians while still maintaining its autonomous functionality. In addition, it highlights the potential for collaboration between human experts and AI systems in the medical field.

As mentioned, this work represents a proof of concept for a fully autonomous simulation framework. Understandably, in future work, we plan to address potential limitations, such as examining potential downfalls, interface design, and considering the mental model of human users. These considerations will contribute to further developing and refining of our framework, making it more robust, user-friendly, and aligned with the needs and expectations of medical professionals.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The study was supported by Technion's TASP grant entitled "Autonomous Medical Simulation and Training".

Appendix A. Supplementary materials

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104446.

References

[1] A. Ziv, P.R. Wolpe, S.D. Small, S. Glick, Simulation-based medical education: an ethical imperative, Simul. Healthcare: J. Soc Simul. Healthc. 1 (4) (2006) 252–256, https://doi.org/10.1097/01.sih.0000242724.08501.63.

[2] D.B. Swanson, G.R. Norman, R.L. Linn, Performance-based assessment: lessons from the health professions, Educ. Res. 24 (5) (1995) 5–11, https://doi.org/10.3102/0013189X024005005.

[3] M. Srinivasan, J.C. Hwang, D. West, P.M. Yellowlees, Assessment of clinical skills using simulator technologies, Acad. Psychiatry 30 (6) (2006) 505–515, https://doi.org/10.1176/appi.ap.30.6.505.

[4] J.R. Boulet, M. Van Zanten, A. De Champlain, R.E. Hawkins, S.J. Peitzman, Checklist content on a standardized patient assessment: An ex post facto review, Adv. Health Sci. Educ. 13 (1) (2008) 59–69, https://doi.org/10.1007/s10459-006-9024-4.

[5] J.R. Boulet, D.J. Murray, Simulation-based assessment in anesthesiology: Requirements for practical implementation, Anesthesiology 112 (4) (2010) 1041–1052, https://doi.org/10.1097/ALN.0b013e3181cea265.

[6] C.J. Gordon, T. Ryall, B. Judd, Simulation-based assessments in health professional education: a systematic review, J. Multidiscip. Healthc. 9 (2016) 69, https://doi.org/10.2147/JMDH.S92695.

[7] J.C. Archer, State of the science in health professional education: Effective feedback, Med Educ. 44 (1) (2010) 101–108, https://doi.org/10.1111/j.1365-2923.2009.03546.x.

[8] B.M. Scavone, M.T. Sproviero, R.J. McCarthy, et al., Development of an objective scoring system for measurement of resident performance on the human patient simulator, Anesthesiology 105 (2) (2006) 260–266, https://doi.org/10.1097/00000542-200608000-00008.

[9] P.J. Morgan, J. Lam-McCulloch, J. Herold-McIlroy, J. Tarshis, Simulation performance checklist generation using the Delphi technique, Can. J. Anesth. 54 (12) (2007) 992–997, https://doi.org/10.1007/BF03016633.

[10] B.G. Shinn-Cunningham, Object-based auditory and visual attention, Trends Cogn Sci. 12 (5) (2008) 182–186, https://doi.org/10.1016/j.tics.2008.02.003.

[11] P.J. Morgan, D. Cleave-Hogg, C.B. Guest, A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator, Acad. Med. 76 (10) (2001) 1053–1055, https://doi.org/10.1097/00001888-200110000-00016.

[12] P. Shayne, F. Gallahue, S. Rinnert, C.L. Anderson, G. Hern, E. Katz, Reliability of a core competency checklist assessment in the emergency department: The standardized direct observation assessment tool, Acad. Emerg. Med. 13 (7) (2006) 727–732, https://doi.org/10.1197/j.aem.2006.01.030.

[13] Z. Zhang, A. Sarcevic, Constructing Awareness Through Speech, Gesture, Gaze and Movement During a Time-Critical Medical Task, in: *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19-23 September 2015, Oslo, Norway*, 2015, pp. 163-182. doi:10.1007/978-3-319-20499-4_9.

[14] S. Gershov, Y. Ringel, E. Dvir, et al. Automatic speech-based checklist for medical simulations, in: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, Association for Computational Linguistics, 2021, pp. 30–34. doi: 10.18653/v1/2021.nlpmc-1.4.

[15] R. Poppe, A survey on vision-based human action recognition, Image Vis Comput. 28 (6) (2010) 976–990, https://doi.org/10.1016/j.imavis.2009.11.014.

[16] E. Kim, S. Helal, D. Cook, Human activity recognition and pattern discovery, IEEE Pervasive Comput. 9 (1) (2010) 48–53, https://doi.org/10.1109/MPRV.2010.7.

[17] M. Vrigkas, C. Nikou, I.A. Kakadiaris, A review of human activity recognition methods, Front. Robot. AI 2 (NOV) (2015) 1–28, https://doi.org/10.3389/frobt.2015.00028.

[18] Ó.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Commun. Surv. Tutorials 15 (3) (2013) 1192–1209, https://doi.org/10.1109/SURV.2012.110112.00192.

[19] D.R. Beddiar, B. Nini, M. Sabokrou, A. Hadid, Vision-based human activity recognition: a survey, Multimed Tools Appl. 79 (41–42) (2020) 30509–30555, https://doi.org/10.1007/s11042-020-09004-3.

[20] S. Ranasinghe, F. Al MacHot, H.C. Mayr, A review on applications of activity recognition systems with regard to performance and evaluation, Int. J. Distrib. Sens. Netw. 12 (8) (2016), https://doi.org/10.1177/1550147716665520.

[21] S. Aarthi, S. Juliet, A comprehensive study on Human Activity Recognition, in: Institute of Electrical and Electronics Engineers Inc, 2021, pp. 59–63, https://doi.org/10.1109/ICSPC51351.2021.9451759.

[22] O.C. Ann, L.B. Theng, Human activity recognition: A review, in: Proceedings - 4th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2014, 2014 (November), pp. 389–393, doi:10.1109/ICCSCE.2014.7072750.

[23] J.M. Chaquet, E.J. Carmona, A. Fernández-Caballero, A survey of video datasets for human action and activity recognition, Comput. Vis. Image Underst. 117 (6) (2013) 633–659, https://doi.org/10.1016/j.cviu.2013.01.013.

[24] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, et al. Multimodal human action recognition in assistive human-robot interaction, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, 2016, 2016-May, pp. 2702–2706, doi: 10.1109/ICASSP.2016.7472168.

[25] J. Abdulbaqi, Y. Gu, Z. Xu, C. Gao, I. Marsic, R.S. Burd, Speech-based activity recognition for trauma resuscitation, in: 2020 IEEE International Conference on Healthcare Informatics, ICHI, 2020, https://doi.org/10.1109/ICHI48887.2020.9374372. Published online 2020.

[26] S. Jagannath, A. Sarcevic, N. Kamireddi, I. Marsic, Assessing the feasibility of speech-based activity recognition in dynamic medical settings, in: Conference on Human Factors in Computing Systems - Proceedings, 2019, https://doi.org/10.1145/3290607.3312983.

[27] R. Gao, T.H. Oh, K. Grauman, L. Torresani, Listen to look: Action recognition by previewing audio, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Published online 2020:10454-10464. doi:10.1109/CVPR42600.2020.01047.

[28] D. Istrate, E. Castelli, M. Vacher, L. Besacier, J.F. Serignat, Information extraction from sound for medical telemonitoring, IEEE Trans. Inf Technol. Biomed. 10 (2) (2006) 264–274, https://doi.org/10.1109/TITB.2005.859889.

[29] N. Weibel, S. Rick, C. Emmenegger, S. Ashfaq, A. Calvitti, Z. Agha, LAB-IN-A-BOX: semi-automatic tracking of activity in the medical office, Pers. Ubiquitous Comput. 19 (2) (2015) 317–334, https://doi.org/10.1007/s00779-014-0821-0.

[30] V. Osmani, S. Balasubramaniam, D. Botvich, Human activity recognition in pervasive health-care: Supporting efficient remote collaboration, J. Netw. Comput. Appl. 31 (4) (2008) 628–655, https://doi.org/10.1016/j.jnca.2007.11.002.

[31] Y. Gu, R. Zhang, X. Zhao, et al., multimodal attention network for trauma activity recognition from spoken language and environmental sound HHS public access, IEEE Int. Conf. Healthc. Inform. (2019), https://doi.org/10.1109/ichi.2019.8904713.

[32] C. Gao, I. Marsic, A. Sarcevic, W. Gestrich-Thompson, R.S. Burd, Real-time context-aware multimodal network for activity and activity-stage recognition from team communication in dynamic clinical settings, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 7 (1) (2023), https://doi.org/10.1145/3580798.

[33] S. Harispe, S. Ranwez, S. Janaqi, J. Montmain, Semantic similarity from natural language and ontology analysis, Synthesis Lect. Human Lang. Technol. 8 (1) (2015) 1–256, https://doi.org/10.2200/S00639ED1V01Y201504HLT027.

[34] Y. Feng, E. Bagheri, F. Ensan, J. Jovanovic, The state of the art in semantic relatedness: A framework for comparison, Knowl. Eng. Rev. (2017) 32, https://doi.org/10.1017/S0269888917000029.

[35] D. Chandrasekaran, V. Mago, Evolution of semantic similarity-A Survey, ACM Comput. Surv. 54 (2) (2021) 41, https://doi.org/10.1145/3440755.

[36] C. Corley, R. Mihalcea, Measuring the semantic similarity of texts, in: EMSEE 2005 - Empirical Modeling of Semantic Equivalence and Entailment@ACL 2005, Proceedings of the Workshop, 2005, pp. 13–18, doi:10.3115/1631862.1631865.

[37] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol 1., 2019, pp. 4171-4186. Accessed April 17, 2022. https://github.com/tensorflow/tensor2tensor.

[38] J.B. Sexton, E.J. Thomas, R.L. Helmreich, Error, stress, and teamwork in medicine and aviation: cross sectional surveys, BMJ 320 (7237) (2000) 745–749, https://doi.org/10.1136/bmj.320.7237.745.

[39] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, An evaluation of outcome from intensive care in major medical centers, Ann. Intern. Med. 104 (3) (1986) 410–418, https://doi.org/10.7326/0003-4819-104-3-410.

[40] S.M. Shortell, J.E. Zimmerman, D.M. Rousseau, et al., The performance of intensive care units: Does good management make a difference? Med. Care. 32 (5) (1994) 508–525, https://doi.org/10.1097/00005650-199405000-00009.

[41] J.P. Brown, Closing the communication loop: using readback/hearback to support patient safety, Jt. Comm. J. Qual. Saf. 30 (8) (2004) 460–464, https://doi.org/10.1016/S1549-3741(04)30053-5.

[42] A.K. Hall, J.D. Dagnone, L. Lacroix, W. Pickett, D.A. Klinger, Queen's simulation assessment tool: Development and validation of an assessment tool for resuscitation objective structured clinical examination stations in emergency medicine, Simul. Healthc. 10 (2) (2015) 98–105, https://doi.org/10.1097/SIH.0000000000000076.

[43] C. Faudeux, A. Tran, A. Dupont, et al., Development of reliable and validated tools to evaluate technical resuscitation skills in a pediatric simulation setting: resuscitation and emergency simulation checklist for assessment in pediatrics, J. Pediatr. 188 (2017) 252–257.e6, https://doi.org/10.1016/j.jpeds.2017.03.055.

[44] T.C. Everett, E. Ng, D. Power, et al., The Managing Emergencies in Paediatric Anaesthesia global rating scale is a reliable tool for simulation-based assessment in pediatric anesthesia crisis management, Paediatr. Anaesth. 23 (12) (2013) 1117–1123, https://doi.org/10.1111/pan.12212.

[45] J. Wallenstein, D. Ander, Objective structured clinical examinations provide valid clinical skills assessment in emergency medicine education, West. J. Emergency Med. 16 (1) (2015) 121–126, https://doi.org/10.5811/westjem.2014.11.22440.

[46] E.C. Cherry, Some experiments on the recognition of speech, with one and with two ears, J. Acoust. Soc. Am. 25 (5) (1953) 975–979, https://doi.org/10.1121/1.1907229.

[47] E. Vincent, T. Virtanen, S. Gannot, in: Audio Source Separation and Speech Enhancement, John Wiley & Sons Ltd, 2018, https://doi.org/10.1002/9781119279860.

[48] Pariente M, Cornell S, Cosentino J, et al. Asteroid: The PyTorch-based audio source separation toolkit for researchers, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol 2020-Octob, 2020, pp. 2637-2641, doi:10.21437/Interspeech.2020-1673.

[49] Manilow E, Seetharaman P, Pardo B. The northwestern university source separation library, in: *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, 2018, pp. 297–305. Accessed April 4, 2022, https://interactiveaudiolab.github.io/demos/nussl.html.

[50] Z. Ni, M.I. Mandel, Onssen: an open-source speech separation and enhancement library. Published online November 3, 2019. Accessed April 4, 2022. https://github.com/speechLabBcCuny/onssen.

[51] Y. Luo, N. Mesgarani, Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (8) (2019) 1256–1266, https://doi.org/10.1109/TASLP.2019.2915167.

[52] O.H. Anidjar, I. Lapidot, C. Hajaj, A. Dvir, I. Gilad, Hybrid speech and text analysis methods for speaker change detection, IEEE/ACM Trans. Audio Speech Lang. Process. 29 (2021) 2324–2338, https://doi.org/10.1109/TASLP.2021.3093817.

[53] T.J. Park, N. Kanda, D. Dimitriadis, K.J. Han, S. Watanabe, S. Narayanan, A review of speaker diarization: Recent advances with deep learning, Comput Speech Lang. (2022) 72, https://doi.org/10.1016/j.csl.2021.101317.

[54] N. Kanda, C. Boeddeker, J. Heitkaemper, et al. Guided source separation meets a strong ASR backend: Hitachi/Paderborn university joint investigation for dinner party ASR, in: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, 2019, pp. 1248–1252. doi:10.21437/Interspeech.2019-1167.

[55] Horiguchi S, Fujita Y, Watanabe S, Xue Y, Nagamatsu K. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, in: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-Octob, 2020, pp. 269–273, doi:10.21437/Interspeech.2020-1022.

[56] T. Giannakopoulos, PyAudioAnalysis: An open-source python library for audio signal analysis, PLoS One 10 (12) (2015) e0144610.

[57] A. Kutuzov, E. Kuzmenko, To lemmatize or not to lemmatize: how word normalisation affects ELMo performance in word sense disambiguation. Published online September 6, 2019. Accessed April 14, 2022. http://arxiv.org/abs/1909.03135.

[58] A. Seker, E. Bandel, D. Bareket, I. Brusilovsky, R.S. Greenfeld, R. Tsarfaty, AlephBERT:A Hebrew Large Pre-Trained Language Model to Start-off your Hebrew NLP Application With. Published online 2021. Accessed April 14, 2022. https://github.com/OnlpLab/AlephBERT/.

[59] R. Tsarfaty, A. Seker, S. Sadde, S. Klein, What's wrong with Hebrew nlp? And how to make it right, in: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations*, 2019, pp. 259–264, doi: 10.18653/v1/d19-3044.

[60] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process Manag. 24 (5) (1988) 513–523, https://doi.org/10.1016/0306-4573(88)90021-0.

[61] M.A. Iqbal, O. Sharif, M.M. Hoque, I.H. Sarkar, Word Embedding based Textual Semantic Similarity Measure in Bengali, in: *Procedia Computer Science*, vol. 193, Elsevier B.V., 2021, pp. 92–101, doi:10.1016/j.procs.2021.10.010.

[62] I. Beltagy, K. Lo, A. Cohan, SCIBERT: A pretrained language model for scientific text, in: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 3615-3620. doi:10.18653/v1/d19-1371.

[63] E. Alsentzer, J.R. Murphy, W. Boag, et al. Publicly Available Clinical BERT Embeddings. Published online 2019. Accessed April 17, 2022. https://www.ncbi.nlm.nih.gov/pmc/.

[64] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference,* Published online 2019, pp. 3982–3992. doi:10.18653/v1/d19-1410.

[65] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Published online 2020, pp. 37-63, http://arxiv.org/abs/2010.16061.

[66] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. 15 (2) (2016) 155–163, https://doi.org/10.1016/j.jcm.2016.02.012.

[67] H.J. Keselman, A.R. Othman, R.R. Wilcox, K. Fradette, The new and improved two-sample t test, Psychol Sci. 15 (1) (2004) 47–51, https://doi.org/10.1111/j.0963-7214.2004.01501008.x.

[68] J. Kim, D. Neilipovitz, P. Cardinal, M. Chiu, J. Clinch, A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, high-fidelity simulation, and crisis resource management I study, Crit. Care Med. 34 (8) (2006) 2167–2174, https://doi.org/10.1097/01.CCM.0000229877.45125.CC.

[69] J.S. Oh, S.G. Kim, S.C. Lim, J.L. Ong, A comparative study of two noninvasive techniques to evaluate implant stability: Periotest and Osstell Mentor, Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endodontol. 107 (4) (2009) 513–518, https://doi.org/10.1016/j.tripleo.2008.08.026.

[70] J.S. Cha, N.E. Anton, T. Mizota, et al., Use of non-technical skills can predict medical student performance in acute care simulated scenarios, Am. J. Surg. 217 (2) (2019) 323–328, https://doi.org/10.1016/j.amjsurg.2018.09.028.

[71] S. Bubeck, V. Chandrasekaran, R. Eldan, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. Published online March 22, 2023. Accessed June 13, 2023. http://arxiv.org/abs/2303.12712.

[72] H. Nori, N. King, S.M. McKinney, D. Carignan, E. Horvitz, Capabilities of GPT-4 on Medical Challenge Problems. Published online March 20, 2023. Accessed June 13, 2023. http://arxiv.org/abs/2303.13375.