Automatic Speech-Based Checklist for Medical Simulations

Sapir Gershov
Technion - Israel Institute of Technology, Haifa, Israel

Dr. Yaniv Ringel, Dr. Erez Dvir, Tzvia Tsirilman, Dr. Elad Ben Zvi, Dr. Sandra Braun, Dr. Aeyal Raz
Rambam Health Care Campus, Haifa, Israel

Dr. Shlomi Laufer
Technion - Israel Institute of Technology, Haifa, Israel
laufer@technion.ac.il

Abstract

Medical simulators provide a controlled environment for training and assessing clinical skills. However, as an assessment platform, it requires the presence of an experienced examiner to provide performance feedback, commonly preformed using a task specific checklist. This makes the assessment process inefficient and expensive. Furthermore, this evaluation method does not provide medical practitioners the opportunity for independent training. Ideally, the process of filling the checklist should be done by a fully-aware objective system, capable of recognizing and monitoring the clinical performance. To this end, we have developed an autonomous and a fully automatic speech-based checklist system, capable of objectively identifying and validating anesthesia residents' actions in a simulation environment. Based on the analyzed results, our system is capable of recognizing most of the tasks in the checklist: F_1 score of 0.77 for all of the tasks, and F_1 score of 0.79 for the verbal tasks. Developing an audio-based system will improve the experience of a wide range of simulation platforms. Furthermore, in the future, this approach may be implemented in the operation room and emergency room. This could facilitate the development of automatic assistive technologies for these domains.

1 Introduction

In recent years, there is a growing interest in developing performance-based assessment for medical practitioners. In the pursuit for methods that may assess hands-on skills, simulation-based assessment has emerged (Srinivasan et al., 2006; Swanson et al., 1995). Simulation-based assessment requires appropriate validation metrics, and checklists are one of the most common methods. For a given simulation scenario, evaluation experts determine which actions, based on the presenting complaint, are important for the candidate to perform in order to properly manage the scenario (Scavone et al., 2006; Morgan et al., 2007). Based on this process, a detailed checklist is developed (Morgan et al., 2007; Hilliard et al., 2000; Morgan et al., 2001; Boulet et al., 2008; Shayne et al., 2006). During the simulation, an experienced examiner is required for filling in the checklist. The need for an experienced examiner makes the assessment process very time consuming and expensive, and in addition, does not provide medical practitioners the opportunity for independent training.

Ideally, to reduce the costs of performance assessments and to allow more residents to train in a complex scenario, the process of filling the checklist should be done by a machine: A fully-aware objective system capable of recognizing and monitoring the resident performance. To this end, we have developed a end-to-end fully automatic speech-based objective checklist validation system, capable of identifying anesthesia residents' actions in a simulation environment, based solely on the participants' speech recordings. We developed a simulation setup for data collecting. The checklist system was evaluated using two different clinical scenarios for assessing skills of senior anesthesia residents. Our underlying assumption is that in many cases the communication among medical staff may represent the physical action itself. By analyzing the participants' speech, our system can automatically identify and fill the appropriate rubrics in the checklist.

2 Materials and Methods

2.1 Medical Simulation

Two clinical scenarios were developed by an experienced anesthesiologist and a medical simulation expert. The scenarios were based on scenarios previously written by the anesthesiologist (A. R) and were used for the Israeli Anesthesiology board certification exam. The first scenario included the management of a patient with a severe anaphy-

Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pages 30–34

laxis reaction and the second scenario involved a patient after surgery suffering from severe bradycardia. The study was approved by the Rambam Medical Center IRB committee.

As done in similar medical simulation studies (Hall et al., 2015; Faudeux et al., 2017; Everett et al., 2013; Wallenstein and Ander, 2015), a detailed checklist was developed for each scenario. The checklist included approximately 35 tasks the participants were expected to perform. The score for each task was in the range of 0-2, representing the performance quality in comparison to standard medical guidelines. The checklist tasks scores are scaled as follows: 0 for not observed, 1 for needs improvement, 2 for meets expectations.

Fifteen senior anesthesiology residents, 11 males and 4 females, participated in the study. Five of them preformed both simulation scenarios, 4 residents preformed only the anaphylaxis scenario and 5 preformed only the bradycardia scenario. In addition, two members of the research team played the roles of a nurse and a medical intern. During the simulation, an experienced anesthesiologist evaluated the resident's performance using the scenario checklist. A 'Laerdal' MegaCode Kelly, a full body manikin designed for the practice of Advanced Cardiovascular Life Support (ACLS), was used as the patient.

Video and audio were recorded using StreamPix digital video recording software (NorPix Inc.). The recorded video data was used by a human observer to manually fill in the checklist. For audio recordings, the resident and the nurse wore a wireless lavalier microphone transmitter (Sony UWP-D11), which was connected to a digital mixer (Tascam US-20x20). Each audio channel was saved separately.

2.2 Automatic Checklist Generation

The automatic generation of the checklist included several steps. First, automatic transcription was performed, and then, keywords were identified in each sentence. Using these keywords, a matching process between the checklist tasks and the corpus sentences was implemented. The outcome of the algorithm was a filled checklist in which the completed tasks are provided with a matching sentence and timestamp. A detailed description of each step will be provided in the following sections (Figure 2).

2.3 Automatic Transcription

The recorded audio data were automatically transcribed using Google's speech-to-Text API. This required two preprocessing steps:

1. Audio Source Separation – since the physician, nurse and intern stood in close proximity, each audio channel recorded multiple speakers as well as background noise (e.g. patient monitors). Thus, the mixed audio signal was separated into individual source signals (Vincent et al., 2018). In recent years, several open-source audio toolkits have provided implementations of audio source separation methods using deep learning (Pariente et al., 2020; Manilow et al., 2018; Ni and Mandel, 2019). In this study we used the Conv-TasNet (Luo and Mesgarani, 2019) network provided by Asteroid (Pariente et al., 2020). The network was fine-tuned for Hebrew speech as well as to our audio recording device.

2. Audio Segmentation – our objective was to provide a transcription with timestamps for each sentence. Therefore, each audio channel was segmented using the 'pyAudioAnalysis' library (Giannakopoulos, 2015). This library provides a semi-supervised audio segmentation using an SVM model. This function takes an uninterrupted audio recording as input and returns segment endpoints that correspond to the areas of "silence" between them. To achieve better division to segments, adjustments of the dynamic thresholds were performed.

2.4 Morphological and Syntactic Parsing

In order to syntactically analyze texts, the input tokens are first broken down to their constituent morphemes. However, Morphologically Rich Language (MRL), such as Hebrew, pose a unique challenge to the standard language processing pipeline. Due to extreme morphological ambiguity, global context is required in order to correctly decompose raw tokens into morphemes (More et al., 2019). To overcome this challenge, a morpho-syntactic parser for morphological and syntactic analysis of Hebrew texts (Tsarfaty et al., 2020) was used. Morphologically rich syntax parsing is useful in cases of verbs and adjectives, by reducing the variance in the transcription database.

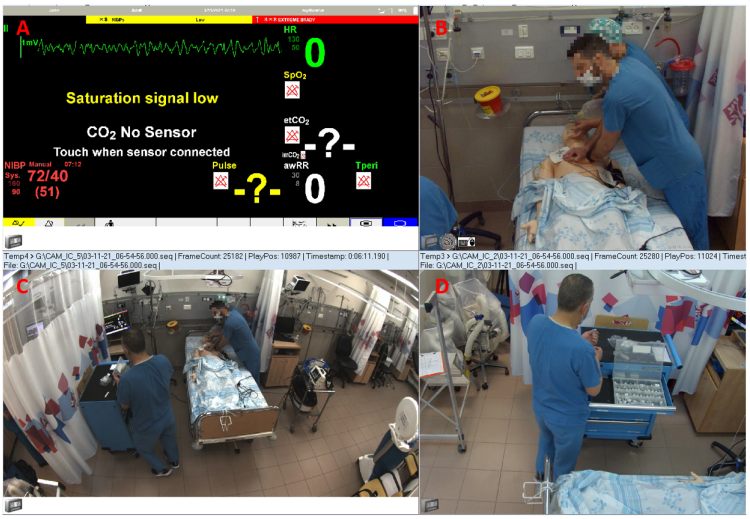

Figure 1: Data acquisition system. (A) Patient monitor; (B) Physician working area; (C) Overview of the simulation area; (D) Nurse working area. In addition, each participant carried a wireless lavalier microphone transmitter

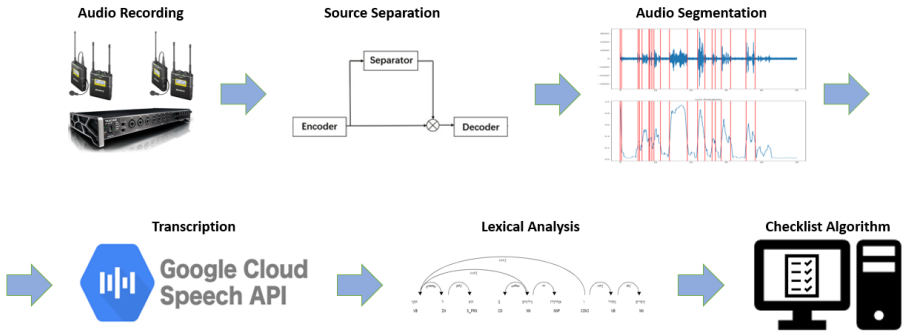

Figure 2: Automatic checklist process. End-to-end description of the checklist generation pipeline.

2.5 Word importance

The checklist includes short descriptions of each task. These descriptions guide the examiner in identifying the different assignments performed by the participants. Hence, by stripping the task description to its base form and choosing distinct words that best represent the task, a bag-of-words for each task can be generated. These "bags" will serve as touchstones for assessing how each sentence in the transcription is suitable to describe the task in hand. These keywords tend to be medical terms (medications, procedures, etc.) and combinations of an object and a verb. The matching process is based on thresholded argmax.

2.6 Checklist Evaluation

After collecting the simulation recording, a professional performance evaluator observed the video recordings and completed the checklist. As mentioned, each task in the checklist received a score in the range of 0-2. However, since the algorithm developed in this study is a binary classier, scores 1 and 2 were considered true (task preformed) and 0 was considered false. The classifier was assessed using the F_1 score (Powers, 2020).

3 Results

As mentioned in section 2.4, a proper syntactically analysis of Hebrew texts requires the disassembling of the input tokens down to their constituent morphemes. To evaluate the impact of the lexical analysis on the algorithm results, we compared two versions of the pipeline - one used the lexical analysis and the other didn't.

During the process of analyzing the data, we found that few tasks in the checklist tend to be non-verbal in their nature. Most candidate don't use any verbal commands when preforming those tasks, and the human observer can validate them only by identifying the action itself. These tasks include 'verification of intubation tube location', 'exposure of patient chest' and few others. These

Algorithm without Lexical analysis				
Division Category	Total tasks	Tasks preformed	Algorithm identified	F_1 score
All	664	405	249	0.682
Verbal	578	363	233	**0.704**
Non-Verbal	86	42	16	0.470

Table 1: Algorithm without lexical analysis performances for all, verbal and non-verbal tasks in the collected data

Algorithm with Lexical analysis				
Division Category	Total tasks	Tasks preformed	Algorithm identified	F_1 score
All	664	405	316	0.773
Verbal	578	363	292	**0.793**
Non-Verbal	86	42	24	0.585

Table 2: Algorithm with lexical analysis performances for all, verbal and non-verbal tasks in the collected data

tasks have a dramatic affect on the performance of our system. Based on this findings, we decided to divide our checklist into two different categories: verbal and non-verbal tasks. This provided us with better understanding of the system limitations. The collected data from the post-simulation human observer and the F_1 scores over all three division: All tasks, Verbal tasks and Non-verbal tasks can be found in Table 1 & 2.

4 Discussion

In this study we developed a system for automatically filling a medical simulation checklist using audio data. The system is completely autonomous and a fully automatic pipeline from the raw audio files to a complete checklist was established. The system was assessed using novel data collected for this study.

The native language of the current participants of this study is Hebrew. This poses a unique challenge common to Morphologically Rich Language. As clearly evident from the results, using lexical analysis improved our system performances, and might have a greater impact on a more complex models. We plan to expand our work to other languages in the future and assess the system performance.

The system was successful in correctly identifying most of the tasks performed by the participants. Yet, one limitation of the system is that it is currently based on keyword matching and not on a more complex model of the conversation. The method in use has limited accuracy, and in addition, only provides a binary score indication whether the task was preformed or not. For example, the current system may indicate a drug was provided but it will not assess the dosage. In order to develop a more complex algorithm, a significantly larger data base is required. We are continuously collecting data that focuses both on a larger number of participants as well as a wide range of clinical scenarios. This will expedite the development of more complex algorithms.

Developing an audio-based system has several advantages. First, it may fit to a wide range of simulation platforms including low- and high-fidelity mannequins, virtual reality, and standardized patients. Furthermore, in the future, our system could migrate from the simulation domain and be implemented in the operation room and emergency room. This could facilitate the development of automatic assistive systems for these domains.

Acknowledgements

The study was supported by the Technion's TASP-2020 grant entitled "Autonomous Medical Simulation and Training".

References

John R. Boulet, Marta Van Zanten, André De Champlain, Richard E. Hawkins, and Steven J. Peitzman. 2008. Checklist content on a standardized patient assessment: An ex post facto review. *Advances in Health Sciences Education*, 13(1):59–69.

Tobias C. Everett, Elaine Ng, Daniel Power, Christopher Marsh, Stephen Tolchard, Anna Shadrina, and Matthew D. Bould. 2013. The Managing Emergencies in Paediatric Anaesthesia global rating scale is a

reliable tool for simulation-based assessment in pediatric anesthesia crisis management.

Camille Faudeux, Antoine Tran, Audrey Dupont, Jonathan Desmontils, Isabelle Montaudié, Jean Bréaud, Marc Braun, Jean Paul Fournier, Etienne Bérard, Noémie Berlengi, Cyril Schweitzer, Hervé Haas, Hervé Caci, Amélie Gatin, and Lisa Giovannini-Chami. 2017. 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. *Journal of Pediatrics*, 188:252–257.

Theodoros Giannakopoulos. 2015. PyAudioAnalysis: An open-source python library for audio signal analysis. *PLoS ONE*, 10(12):1–17.

Andrew Koch Hall, Jeffrey Damon Dagnone, Lauren Lacroix, William Pickett, and Don Albert Klinger. 2015. Queen's simulation assessment tool: Development and validation of an assessment tool for resuscitation objective structured clinical examination stations in emergency medicine.

Robert I. Hilliard, Susan E. Tallett, and Diana Tabak. 2000. Use of an Objective Structured Clinical Examination as a Certifying Examination in pediatrics.

Yi Luo and Nima Mesgarani. 2019. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 27(8):1256–1266.

Ethan Manilow, Prem Seetharaman, and Bryan Pardo. 2018. The northwestern university source separation library. *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, pages 297–305.

Amir More, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint Transition-Based Models for Morpho-Syntactic Parsing: Parsing Strategies for MRLs and a Case Study from Modern Hebrew. *Transactions of the Association for Computational Linguistics*, 7(2001):33–48.

Pamela J. Morgan, Doreen Cleave-Hogg, and Cameron B. Guest. 2001. A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. *Academic Medicine*, 76(10):1053–1055.

Pamela J. Morgan, Jenny Lam-McCulloch, Jodi Herold-McIlroy, and Jordan Tarshis. 2007. Simulation performance checklist generation using the Delphi technique. *Canadian Journal of Anesthesia*, 54(12):992–997.

Zhaoheng Ni and Michael I. Mandel. 2019. Onssen: an Open-Source Speech Separation and Enhancement Library. *arXiv*.

Manuel Pariente, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian Robert Stöter, Mathieu Hu, Juan M. Martín-Doñas, David Ditter, Ariel Frank, Antoine Deleforge, and Emmanuel Vincent. 2020. Asteroid: The PyTorch-based audio source separation toolkit for researchers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020-Octob(1):2637–2641.

David M. W. Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. pages 37–63.

Barbara M. Scavone, Michele T. Sproviero, Robert J. McCarthy, Cynthia A. Wong, John T. Sullivan, Viva J. Siddall, and Leonard D. Wade. 2006. Development of an objective scoring system for measurement of resident performance on the human patient simulator. *Anesthesiology*, 105(2):260–266.

Philip Shayne, Fiona Gallahue, Stephan Rinnert, Craig L. Anderson, Gene Hern, and Eric Katz. 2006. Reliability of a Core Competency Checklist Assessment in the Emergency Department: The Standardized Direct Observation Assessment Tool. *Academic Emergency Medicine*, 13(7):727–732.

Malathi Srinivasan, Judith C. Hwang, Daniel West, and Peter M. Yellowlees. 2006. Assessment of clinical skills using simulator technologies.

David B. Swanson, Geoffrey R. Norman, and Robert L. Linn. 1995. Performance-Based Assessment: Lessons From the Health Professions. *Educational Researcher*, 24(5):5–11.

Reut Tsarfaty, Amit Seker, Shoval Sadde, and Stav Klein. 2020. *What's wrong with Hebrew nlp? And how to make it right*.

Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot, editors. 2018. *Audio Source Separation and Speech Enhancement*. John Wiley & Sons Ltd, Chichester, UK.

Joshua Wallenstein and Douglas Ander. 2015. Objective structured clinical examinations provide valid clinical skills assessment in emergency medicine education.