Assessing the Feasibility of Speech-Based Activity Recognition in Dynamic Medical Settings

Swathi Jagannath[1], Aleksandra Sarcevic[1], Neha Kamireddi[1], Ivan Marsic[2]

[1]Drexel University
Philadelphia, PA, USA
{sj532, aleksarc, nk593}@drexel.edu

[2]Rutgers University
Piscataway, NJ, USA
marsic@rutgers.edu

ABSTRACT

We describe an experiment conducted with three domain experts to understand how well they can recognize types and performance stages of activities using speech data transcribed from verbal communications during dynamic medical teamwork. The insights gained from this experiment will inform the design of an automatic activity recognition system to alert medical teams to process deviations in real time. We contribute to the literature by (1) characterizing how domain experts perceive the dynamics of activity-related speech, and (2) identifying the challenges associated with system design for speech-based activity recognition in complex team-based work settings.

1 INTRODUCTION

Dynamic medical scenarios such as trauma or emergency medical resuscitations are team-based processes that focus on the initial evaluation and management of severely injured or sick patients in the emergency department (ED). A typical resuscitation team consists of an attending surgeon or ED physician, a surgical fellow or a senior resident, a medication nurse, a scribe nurse, two or three bedside

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5971-9/19/05.
DOI: https://doi.org/10.1145/3290607.3312983

CCS CONCEPTS

• Human-centered computing → Activity centered design • Computing methodologies → Activity recognition and understanding

KEYWORDS

Speech analysis; speech modeling; narrative schema; activity recognition; decision support; emergency medicine

ACM Reference format:

Swathi Jagannath, Aleksandra Sarcevic, Neha Kamireddi, and Ivan Marsic. 2019. Assessing the Feasibility of Speech-Based Activity Recognition in Dynamic Medical Settings. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. 6 pages. https://doi.org/10.1145/3290607.3312983

nurses, an anesthesiologist and a respiratory therapist. Although teams perform resuscitations based on protocols, such as Advanced Trauma Life Support (ATLS) [1], errors and process deviations are common, even among experienced teams [3]. Delays as little as five seconds can lead to 20% decrease in patient survival in the context of life-threatening injuries [6]. Our long-term research goal is to develop a decision support system that relies on multiple sensor modalities (e.g., Radio-Frequency Identification (RFID) technology, computer vision, speech recognition, and other sensors) to detect and recognize team activities, and then alert teams to errors and process delays in real time. In this work, we assess the feasibility of speech as a modality for automatic activity recognition by determining how well a human medical expert can recognize activity type and performance stage (preparation, execution, assessment) based on speech data only. Characterizing how human experts perceive and understand the dynamics of speech, as well as what challenges they face in recognizing activities, will allow us to derive the guidelines for designing a speech-based automatic activity recognition system.

1.1 Background, Related Work and Research Questions

Information sharing through verbal communication during emergency medical resuscitations is critical for establishing a shared mental model of the activity progress and completions [12]. Speech is also important for detecting activity stages because many activities are performed without objects (e.g., by palpation) but their progress is verbally reported. Speech could, therefore, be used as a robust cue for activity type and stage recognition. In our prior work [7], we identified verbal communication patterns during trauma resuscitation, finding that trauma teams use domain-specific phrases and keywords. The challenge, however, is that team members usually do not name the activities while performing them. Rather, teams discuss the plans for an activity (preparation stage), report the progress (execution stage), or report the activity results (assessment stage). Understanding the nuances of these communication patterns and how humans perceive them can provide valuable insights for designing a speech-based activity recognition system.

Several prior studies explored activity recognition in complex team settings by detecting user location via sensors [2], video [4], computer vision [10], and RFID [9]. Some studies have also examined audio-based activity recognition in daily life [5] and clinical settings [8]. These studies, however, have not addressed the challenges of speech-based activity recognition in dynamic medical teamwork. We contribute to this body of work by addressing three research questions: (1) How well a domain expert can recognize activity types and performance stages using a single ("most recent") speech sentence? (2) How does the recognition accuracy change when the expert is allowed to correct past predictions and what are the properties of sentences that trigger the change? (3) What are the challenges in recognizing activities by using only speech data? To answer these questions, we ran an experiment with three experts in trauma resuscitation. We found that the accuracy for activity type recognition was 85% and for activity stage recognition 84%. After the experts were allowed to make corrections for past predictions based on the most recent sentence, their accuracy for activity type and stage recognition rose to 87%. The challenges included the inconsistency between activity performance and verbal reports, succinct and non-grammatical nature of verbal communication, and overlapping

Time	Speech	Recognized Activity Type	Recognized Activity Stage
0:15:40	Equal breath sounds bilateral	Chest Auscultation	assessment
0:15:43	Pulses?	Pulse Check	preparation
0:15:53	Alright do a quick GCS.	GCS Calculation	preparation
0:16:10	124 over 80 Manually	Manual BP Check	assessment

(a)

Ground Truth Activity Log			Corresponding Lines from Transcript	
Activity Type	**Start Time**	**End Time**	**Speech Time**	**Recognized Activity Stage**
Chest Auscultation	**00:15:22**	**00:15:40**	**0:15:40**	**assessment**
Pulse Oximetry	00:15:23	00:15:24		
Passive Oxygen Applied	00:15:28	00:15:29		
Oxygen Held	00:15:28	00:15:29		
Manual BP Check	**00:15:39**	**00:16:08**	**0:16:10**	**assessment**
Pulse Check	**00:15:45**	**00:15:48**	**0:15:43**	**preparation**
Heart Rate	00:16:04	00:16:05		
Right Pupil Check	00:16:07	00:16:11		
Left Pupil Check	00:16:12	00:16:17		
Temperature Check	00:16:36	00:16:54		
GCS Calculation	**00:16:52**	**00:16:54**	**0:15:53**	**preparation**
Automatic BP Check	00:17:04	00:17:05		

(b)

Figure 1: (a) Excerpt from a transcript with the activity type and stage predicted by a participant (Case ID: 160621, Participant ID: 3). (b) Excerpt from the corresponding activity log with correlated lines from transcript. Activities with speech lines from the above transcript are shown in bold.

multi-person speech and interleaved activities. Despite these challenges, the experts were able to derive context from speech lines and effectively perform activity recognition.

2 METHODS

2.1 Dataset

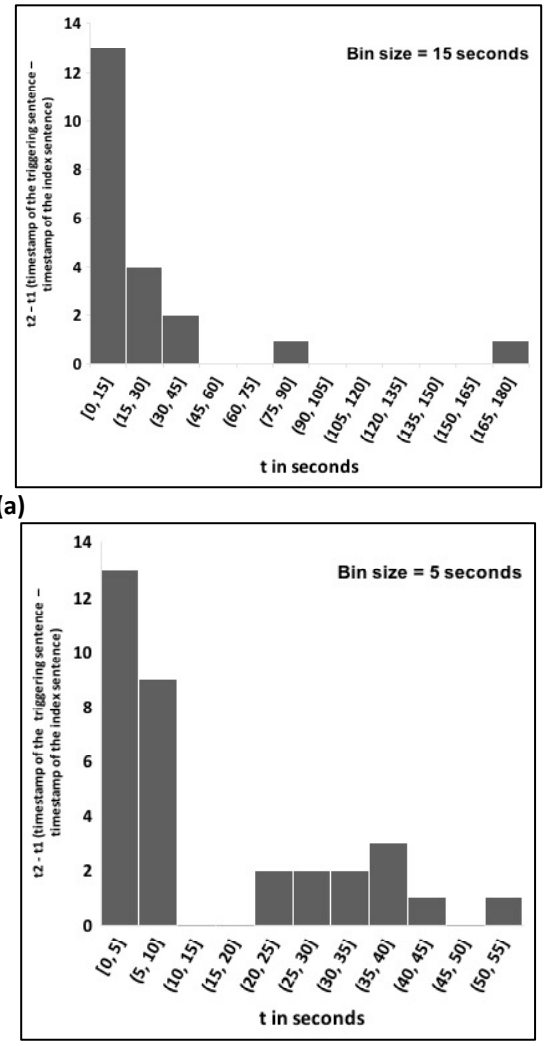
We randomly selected five out of 11 resuscitations performed over three months (June-August, 2016) at a pediatric teaching hospital in the U.S. Mid-Atlantic region and manually transcribed them from microphone recordings. The transcripts included all speech utterances and their timestamps in chronological order. On average, the transcripts had 90 lines (SD = ±24). An activity log derived through video review by three medical experts on our team served as the ground truth data. This activity log contained start and end times for every activity performed during the five resuscitations. The experts were trained in coding team performance on a sample of resuscitations, and proceeded with video annotation only after their inter-rater reliability achieved a Kappa value of >.80 when compared to experienced coders on the team. The experts also developed a data dictionary that defines all activities based on the ATLS protocol.

2.2 Experiment Design

The participants included two ED fellows and an ED nurse, all with multi-year experience in trauma resuscitation. Using electronic chat (e.g., Skype), each participant was shown speech sentences one-by-one from a given transcript, and asked to predict the activity type and stage for the most recently shown sentence. The participants provided their best predictions for the activity type (e.g., Blood Pressure Check, Verbal Airway Assessment) and activity stage (preparation, execution, assessment) for every sentence in all five transcripts. The participants were allowed to reference the data dictionary while predicting activity types and stages. We also allowed them to change their past predictions at any time during the experiment. The past sentence for which the previous prediction is modified is called "index" sentence. The sentence that triggered the modification is called "triggering" sentence. We probed the participants with contextual inquiries when they changed their past predictions to understand their rationale. For the purposes of this experiment, we ignored utterances such as "*ok*" or "*thank you*" since they did not provide valuable information about an activity. The experiment was conducted in two to three sessions with each participant. The sessions ranged from 1.5 to 4.5 hours, for a total of 28 hours. Participants completed one to two transcripts per session, depending on their availability.

2.3 Data Analysis

We performed a five-step data analysis. In *step one,* we aligned timestamps from transcripts with those from the ground truth activity log by using patient arrival time as the reference point. In *step two*, we transferred the participants' chat responses to the transcripts and correlated them with the actual team activity performances from the ground truth data. For example, one participant correctly predicted the speech line at 00:15:40 as the activity type *Chest Auscultation* and activity stage *assessment* (Figure



(a)

(b)

Figure 2: Histograms showing the time difference between the index sentence with the participants' initial prediction and the triggering sentence when they changed their responses. a) *Activity type* prediction changes (b) *Activity stage* prediction changes.

1(a)). As shown in the activity log (Figure 1(b)), the speech-line time correlated with the activity performance between 00:15:22 and 00:15:40 (Chest Auscultation). We excluded 7% of the recognized sentences because we did not have their corresponding ground truth data. In *step three*, we used three heuristics to determine the accuracy of the activity stage recognition: (a) if the speech-line time occurs before the start of activity performance, the speech line indicates *preparation*, (b) if the speech-line time occurs between the start and end of activity performance, the speech line indicates *execution*, and (c) if the speech-line time occurs after the end of activity performance, the speech line indicates *assessment.* In *step four*, we analyzed the content of correction-triggering sentences and their temporal distribution relative to index sentences. In the final *step five*, we identified the challenges of activity recognition based on participant responses to contextual inquiries and patterns of incorrect responses.

3 FINDINGS & DISCUSSION

The participants predicted the activity type and stage for 82% (1110/1356) of sentences but could not classify the remaining sentences for two reasons: (1) sentence was incomplete or contained insufficient information for activity recognition (e.g., *"Ok, turn it off," "No obvious trauma to [unintelligible]"*) and (2) sentence was not about any activity (e.g., *"It's ok sweetie, you can relax"*). Some sentences, however, were associated with more than one activity (e.g., *"Temp 36.8 and BP 124 over 80"*), resulting in more predictions than the number of sentences. The achieved accuracy for activity type was 85% and for activity stage 84%, on average. The accuracy for both activity type and stage recognition rose to 87% after the participants were allowed to change their original predictions using the most recent sentence. The participants changed their responses in 5% cases (54/1156 total predictions), of which 89% (48/54) were correct predictions rectifying the original ones. Of the remaining 11% cases, in 3.7% (2/54) the activity type/stage originally was not predicted when the index sentence was shown, but the subsequent prediction based on a triggering sentence was wrong. In 3.7% (2/54) cases, the original correct prediction was turned to a wrong one. Finally, in 3.7% (2/54) cases, the original wrong prediction was changed to an activity type/stage that was wrong as well.

On average, participants changed their predictions for activity type 25 seconds post index sentence, and for activity stage 15 seconds post index sentence, with a maximum of 2 minutes 51 seconds after the index sentence was uttered. Ninety percent (19/21) changes occurred within 45 seconds for activity type and 67% (22/33) changes occurred within 10 seconds for activity stage (Figure 2). These insights about the temporal distribution of prediction modifications will be useful for designers to understand how long the system may need to wait before making an accurate prediction. Although resuscitation process is fast paced, these delay ranges are acceptable in the few cases when activity recognition based only on individual sentences is incorrect or ambiguous.

The improvements in accuracy showed that the activity recognition depends not only on the most recent ("current") sentence, but also on the sentences that precede or follow it, thus providing better context. For example, the index sentence, *"Can you open your mouth?"* was initially recognized as activity type *Mouth Visual Inspection* and activity stage *preparation*. After the correction-triggering sentence *"Seal your lips around the thermometer"* was shown, the response was changed to activity type *Temperature Check* and activity stage *preparation*. As the participants were seeing successive

Table 1: Excerpts from conversations between the researcher and participants during the experiment.

Excerpt a: Participant 2: *"I had to scroll up to check if they were still in the same activity."*
Excerpt b: Participant 3: *"Could I come back to this line later? I want to see where this is going."*
Excerpt c: Participant 1: *"I lost track on the progress of activities because they are talking about so many things at once. I am scrolling back and forth to make sure I am doing this right."*

sentences, they were creating a narrative of the resuscitation, keeping track of all activities and their progress to assist them with the activity recognition. Our analysis of verbalizations of participants' thought processes during the experiment shows how they were forming the contextual knowledge about activities from the utterances (Table 1). Based on the analysis of these verbalizations and the content of correction-triggering sentences, we concluded that participants used (1) the "topic" of the current sentence and (2) their knowledge of the process workflow to make a modified prediction for the index sentence. In 54% (29/54) cases, both the index and triggering sentences were on the same "topic," i.e., about the same or similar activity. In the remaining 46% (25/54) cases, the triggering sentence was about an activity that reliably follows (in the workflow) the activity to which the index sentence referred. We also found that the on-topic triggering sentences appeared closer to the index sentences, while workflow-related triggering sentences appeared further away from the index sentence. We next describe the challenges with activity recognition based on speech data only:

Inconsistency between activity performance and verbalizations: Due to the dynamic and noisy nature of the resuscitation setting, team members often request for repeating the results of a previously performed activity. This inconsistency between the actual activity performance and the verbalizations that surround it poses a challenge for activity stage recognition because the progress and completion of the activity may not be reflected in verbal report. For example, the speech line *"Did we get a temperature?"* was recognized as a preparatory stage for the Temperature Check activity. Although the speech line implies that the activity is not performed yet, this particular line was a request to confirm the results of the activity that was performed earlier.

Succinct and non-grammatical nature of verbal communication: Although team members use domain-specific keywords and phrases to communicate activity-related information [7], utterances are mostly succinct and grammatically incomplete, leading to challenges for activity recognition. Missing words such as verbs posed the biggest challenge. For example, one participant predicted the speech line *"L R? Yes 1 liter"* as the assessment stage for the administering intravenous fluids activity. The ground truth data log, however, showed that this speech line occurred during the activity execution. In addition, speech lines associated with the same activity prior to this index sentence were missing from transcripts because they were not spoken or unintelligible and could not be transcribed. These missing sentences resulted in a lack of verbal context for activity stage prediction.

Overlapping and interleaving activities: Overlapping multi-person speech and interleaved activities during resuscitations also posed a challenge in understanding the context, keeping track of each activity, and correctly predicting activity stages. Because some activities are discussed and planned, but not performed, the speech is often disconnected. Missing lines from the transcript exacerbated this barrier, resulting in further loss of context and inaccurate activity stage predictions.

4 CONCLUSION

Our findings showed that domain experts can effectively recognize activity type and stage by using speech data only, suggesting that speech recognition is a feasible modality for activity recognition. Despite the challenges, the participants successfully established connections between the activity performances and verbalizations, using the contextual cues to help with activity recognition. They used

ACKNOWLEDGMENTS

This research has been supported by the NSF Awards, Numbers 1253285 and 1763509, as well as by the National Institutes of Health under Award Number R01LM011834. We thank the medical staff at the hospital for their participation.

both past and future sentences in relation to the index sentences as context for their predictions. Finally, we found that the participants associated some speech lines with an activity, but ground truth data from video review were not available. This result has suggested that, in some cases, speech may even provide stronger cues for activity prediction than other modalities (i.e., video). While other sensor modalities, such as computer vision or RFID, could detect activity performance, they usually cannot detect the activity stages of preparation or assessment. Our study has shown that speech could help determine whether the activity was only considered or also performed. Because not all sentences or words will contribute to activity recognition, the system can be designed by training an attention model [11] to learn the impact of words and sentences. We will continue exploring how the challenges associated with speech-based activity recognition could be overcome by combining other sensors in the environment, such as video, computer vision and RFID as part of a decision-support system for complex medical teamwork.

REFERENCES

[1] American College of Surgeons, Advanced Trauma Life Support® (ATLS®), 7th Edition, Chicago, IL, 2005.

[2] Jakob E. Bardram, Afsaneh Doryab, Rune M. Jensen, Poul M. Lange, Kristian L.G. Nielsen, and Søren T. Petersen. 2011. Phase recognition during surgical procedures using embedded and body-worn sensors. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference*, IEEE, 45-53.

[3] Elizabeth A. Carter, Lauren J. Waterhouse, Mark L. Kovler, Jennifer Fritzeen, and Randall S. Burd. 2013. Adherence to ATLS primary and secondary surveys during pediatric trauma resuscitation. *Resuscitation* 84, 1 (Jan. 2013), 66-71.

[4] Ishani Chakraborty, Ahmed Elgammal, and Randall Burd. 2013. Video based activity recognition in trauma resuscitation. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, 1-8.

[5] Theodoros Giannakopoulos and Georgios Siantikos, 2016. A ROS framework for audio-based activity recognition. In *Proc. 9th ACM Int'l Conf. Pervasive Technologies Related to Assistive Environments*. ACM, New York, NY, USA, 41.

[6] Russell L. Gruen, Gregory J. Jurkovich, Lisa K. McIntyre, Hugh M. Foy, and Ronald V. Maier. 2006. Patterns of errors contributing to trauma mortality: lessons learned from 2594 deaths. *Annals of Surgery* 244, 3 (Sep. 2006), 371.

[7] Swathi Jagannath, Aleksandra Sarcevic, and Ivan Marsic. 2018. An analysis of speech as a modality for activity recognition during complex medical teamwork. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ACM, New York, NY, USA, 88-97.

[8] Michael Kranzfelder, Armin Schneider, Sonja Gillen, and Hubertus Feussner. 2011. New technologies for information retrieval to achieve situational awareness and higher patient safety in the surgical operating room: the MRI institutional approach and review of the literature. *Surgical Endoscopy* 25, 3 (Mar. 2011), 696-705.

[9] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall Burd. 2016. Deep learning for RFID-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, New York, NY, USA, 164-175.

[10] Xinyu Li, Yanyi Zhang, Jianyu Zhang, Yueyang Chen, Huangcan Li, Ivan Marsic, and Randall Burd. 2017. Region-based activity recognition using conditional GAN. In *Proceedings of the 2017 ACM on Multimedia Conference*, ACM, New York, NY, USA, 1059-1067.

[11] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 3104-3112.

[12] Zhan Zhang, and Aleksandra Sarcevic. 2015. Constructing awareness through speech, gesture, gaze and movement during a time-critical medical task. In *Proceedings of the 14th European Conference on Computer Supported Cooperative Work, 19-23 September 2015, Oslo, Norway*, Springer, Cham, 163-182.