Real-time Context-Aware Multimodal Network for Activity and Activity-Stage Recognition from Team Communication in Dynamic Clinical Settings

CHENYANG GAO, Rutgers University, United States

IVAN MARSIC, Rutgers University, United States

ALEKSANDRA SARCEVIC, Drexel University, United States

WAVERLY GESTRICH-THOMPSON, Children's National Medical Center, United States

RANDALL S. BURD, Children's National Medical Center, United States

In clinical settings, most automatic recognition systems use visual or sensory data to recognize activities. These systems cannot recognize activities that rely on verbal assessment, lack visual cues, or do not use medical devices. We examined speech-based activity and activity-stage recognition in a clinical domain, making the following contributions. (1) We collected a high-quality dataset representing common activities and activity stages during actual trauma resuscitation events–the initial evaluation and treatment of critically injured patients. (2) We introduced a novel multimodal network based on audio signal and a set of keywords that does not require a high-performing automatic speech recognition (ASR) engine. (3) We designed novel contextual modules to capture dynamic dependencies in team conversations about activities and stages during a complex workflow. (4) We introduced a data augmentation method, which simulates team communication by combining selected utterances and their audio clips, and showed that this method contributed to performance improvement in our data-limited scenario. In offline experiments, our proposed context-aware multimodal model achieved F_1-scores of 73.2±0.8% and 78.1±1.1% for activity and activity-stage recognition, respectively. In online experiments, the performance declined about 10% for both recognition types when using utterance-level segmentation of the ASR output. The performance declined about 15% when we omitted the utterance-level segmentation. Our experiments showed the feasibility of speech-based activity and activity-stage recognition during dynamic clinical events.

CCS CONCEPTS: • **Computing methodologies → Machine learning** • Applied computing → Life and medical sciences → Health care information systems.

Additional Keywords and Phrases: activity recognition, activity-stage recognition, keyword spotting, context-aware recognition, real-time application

Authors' addresses: Chenyang Gao, cg694@rutgers.edu, Rutgers University, Piscataway, United States; Ivan Marsic, marsic@rutgers.edu, Rutgers University, Piscataway, United States; Aleksandra Sarcevic, aleksarc@drexel.edu, Drexel University, Philadelphia, United States; Waverly Gestrich-Thompson, wgestricht@childrensnational.org, Children's National Medical Center, Washington, United States; Randall S. Burd, rburd@childrensnational.org, Children's National Medical Center, Washington, United States.

2474-9567/2023/3–ART12 $15.00
https://doi.org/10.1145/3580798

ACM Reference Format:
Chenyang Gao, Ivan Marsic, Aleksandra Sarcevic, Waverly Gestrich-Thompson, and Randall S. Burd. 2023. 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, Article 12 (March 2023), 28 pages. https://doi.org/10.1145/3580798

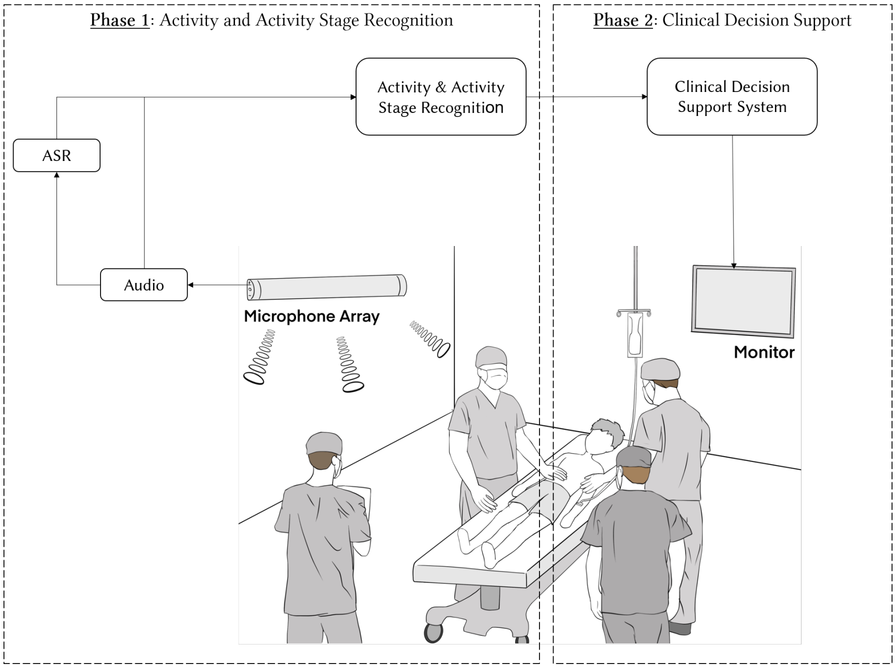
Phase 1: Activity and Activity Stage Recognition
Phase 2: Clinical Decision Support
Activity & Activity Stage Recognition
Clinical Decision Support System
ASR
Audio
Microphone Array
Monitor

Fig. 1: The overall framework of clinical decision support system. This paper focuses on **Phase 1**.

1 INTRODUCTION

Clinical decision support systems (CDSS) [65, 69] can help reduce errors in complex medical teamwork [19]. For CDSS that relies on current task performance as an input, timely and accurate activity recognition is critical. Most activity recognition approaches in medical domains use visual and sensory data and cannot recognize activities that are reported through speech, lack distinct visual cues, or do not use medical devices. However, speech and team communication are critical components of medical work. Medical teams use speech to assign an activity, report the progress or results of the activity, and plan the subsequent activities. The awareness of activities and their temporal progression (activity stage) [30] will affect the next-step decision in clinical decision support systems [38, 65]. Gu et al. [24] proposed a multimodal network that predicts activities from the manually-derived transcripts and audio. This system requires high-performing automatic speech recognition (ASR) to replace manual transcripts, a feature limiting implementation in the real-time because the performance of ASR is inadequate due to the extreme background noise. To address this limitation, Abdulbaqi et al. [1] simplified the recognition problem by using a keyword-based approach, training on a limited-size vocabulary and relying on the identification of the most frequent keyword associated with each activity. Even when a given keyword is associated with more than one activity, this system always infers the activity most frequently associated with that keyword. Another approach relied on speech intention recognition [22] to distinguish semantic differences in utterances spoken during trauma resuscitation events. Although these previous approaches [1, 22, 24] have shown promising results in the recognition of activities and speech intentions, they are inadequate for real-time use and decision support for several reasons. First, the number of previously recognized activity labels has been small (between 5-10) compared to >100 different activity types potentially performed during most clinical events. Second, speech intentions representing the intent of utterances have been too fine-grained with many classes that were not relevant to clinical decision support, and were not always linked to activities, which makes them inadequate for representing the activity progression. Third, they needed manual transcription of human speech which hindered real-time application. Fourth, the conversational context has not been considered. Lastly, predictions were made using utterance-level data, but utterance-level segmentation is not available for real-time use.

We developed a novel context-aware multimodal network to address the aforementioned limitations and bring our model closer to real-time application (Fig. 1). First, we extended the number of activities to include 23 commonly performed activities during trauma resuscitation. We also identified four distinct activity-stage labels from speech intentions [22] based on previously defined features representing activity progression [30-32] to supplement a clinical decision support system [65, 69]. Second, we extended the keyword-based method [1] by using a selected keyword list rather than entire vocabulary. The keyword-based approach avoids the need for a high-performing ASR in real-time application. Third, we designed a novel multimodal network that relies on the conversational context during trauma resuscitation to recognize contextual dependencies of activities and activity stages. Finally, we augmented our dataset by simulating multi-label utterances (i.e., several activities appearing in one utterance) to improve the model's ability to manage concurrent activities and activity stages and avoid or manual segmentation of utterances. We ran experiments with our model in offline (using manually segmented and transcribed speech) and online (using real-time speech recognition without manual segmentation) settings. The results showed a consistent performance improvement with our proposed context modules and data augmentation methods, while also showing the applicability of an offline-trained model in the online setting. In offline and online settings, our model outperformed the NLP (Natural Language Processing) and multimodal baselines that were constructed using existing techniques [4, 17, 28]. We have developed our system using the activities and activity stages specific to trauma resuscitation, but the system can be applied to other settings where activity recognition is also based on verbal communication and the activities are temporally dependent. For example, our system can be applied to settings such as operating rooms and intensive care units (ICU) where many routine activities are also performed (e.g., airway and breathing assessment, oxygen administration, and neurological exam), and the vocabulary used to communicate these activities is the same across these settings. As another example, in daily living activity recognition [2, 44], the activity "washing dishes" is likely to happen after "cooking," and "shaving" usually is followed by "washing face." We believed that modeling the temporal relations between activities in general could improve recognition performance. In this way, our system is generalizable because these temporal dependencies are common across many use cases.

This work makes five contributions to speech-based recognition and monitoring of collocated, dynamic teamwork:

1) A medical dataset with a large set of activity labels that cover commonly occurring activities and their stages during a complex workflow.

2) Approaches that do not require a large vocabulary or a high-performing ASR, including keyword spotting and selection methods that rely on a small set of keywords for activity and activity-stage recognition.

3) An end-to-end multimodal architecture that uses speech sound, keywords, and the conversational context and outperforms previous models that lack the context of team communication.

4) Generation of a simulated team-communication dataset that contains concurrent activities and activity stages that improved the model's ability to recognize concurrent activities and activity stages in data-limited scenarios.

5) Evaluation and discussion of our model in offline experiments and during real-time application.

2 BACKGROUND AND RELATED WORK

The initial care of severely injured patients in the hospital ("trauma resuscitation") is performed using a defined sequence of evaluation and treatment steps. Prior work has found that efficient and error-free care during trauma resuscitation lowers morbidity and mortality [21, 27, 62]. To limit the impact of human factors in this domain, Advanced Trauma Life Support (ATLS) was developed as a standard resuscitation protocol [59]. Deviations from and errors in application of ATLS are still observed, even among experienced trauma teams [20, 48]. Some errors may have no direct impact on patient care, but others can lead to poor outcomes, including death [21]. Given the evidence supporting the benefits of adherence to ATLS, previous studies have evaluated real-time decision support to improve compliance with standard protocols [19]. Although promising, these early systems have had limited usability in practice because they required manual data entry and active interaction with the system. Activity recognition has been proposed for the resuscitation domain to address these limitations by automating data input.

Recognizing human activity is a challenging problem [33]. Different types of activities are defined based on the application needs. In clinical settings, recognition is targeted at critical activities performed by providers during patient care. Recognition of these activities can be used to identify the phases of a clinical process [69], define differences in activity sequence performance associated with different outcomes, and track work in runtime for providing activity-based decision support [65, 69]. Most prior studies have focused on activity recognition in the operating room [55, 58]. Unlike this setting, trauma resuscitation follows a more flexible medical workflow under time pressure, with patient information emerging throughout the process. Activity recognition in team-based settings like trauma resuscitation is challenging for many reasons, including a crowded workspace, concurrent and fast-paced activities, and different modalities for conveying information (e.g., visual observation, verbal assessment, and signals from devices and instruments). Current research on activity recognition mostly relies on visual and sensory data sources to predict activities [5, 10, 42]. Visible-light or depth cameras provide rich sources of data about activities without interfering with work. Object localization in video frames [43] can also assist with activity recognition. This approach has been used in a multimodal network developed to detect the workflow phase during trauma resuscitation [41]. Vision-based approaches have several limitations, including privacy issues, occlusion in a crowded workspace, and the visual similarity of some activities. Unlike vision-based systems, RFID-based recognition systems rely on tagging medical objects and tracking their movement for activity recognition [40, 42]. Passive RFID cannot address challenges of the limited types of activities that use taggable devices and the need for continuous tagging of disposable objects.

In addition to using vision- and sensor-based systems, audio or speech signals can also serve as a valuable information source for human activity recognition. Nicholas et al. [36], pretrained a Gaussian Restricted Boltzmann machine (RBM) with a large set of unannotated audio streams using unsupervised learning. Their model was fine-tuned to various acoustic classification tasks and showed robustness to a wide range of acoustic scenes based on mobile audio sensing. Yohan et al. [14] used multi-sensory data, including speech words and scene sounds, to categorize different places. Dawei et al. [44] used a large-scale pretrained

VGGish model as a feature extractor to generate acoustic embeddings for ambient sounds, followed by a classification network, to perform daily living activity recognition and showed promising results. Rebecca et al. [2] evaluated the feasibility of using mid-interaction segments to distinguish daily living activities, a setting in which mid-interaction segments refer to voice assistants' response time after receiving human queries. A multimodal system was built that recognizes concurrent activities during trauma resuscitation using multiple data modalities, including depth camera video, RFID sensors, and audio recordings [12]. A separate convolutional neural network (CNN) was used to extract features for each modality that were fused using a long short-term memory (LSTM) network in the final decision layer. These previous studies showed the feasibility of using audio or speech related to activity recognition. In clinical settings, vision- and sensor-based systems are not applicable to recognizing activities that lack distinct visual cues or do not use medical devices, such as those performed or reported through verbal communication. For example, airway assessment is performed by asking the patient to answer several questions and verbalizing the assessment results to the entire team. Verbal communication associated with these activities has many cues about the activity and can serve as a valuable source of data for activity recognition in clinical settings [32].

Recent research on predicting activities from manual speech transcripts and audio recordings [24] showed promising results, despite the ambient noise and concurrent speech affecting system performance. In this previous work, a multimodal transformer network manually processed the transcribed speech and audio sound to predict activities. When using only audio, the system predicted the activities with an average accuracy of 36.4%. The accuracy increased to 71.8% for 11 activities when using both modalities (audio recordings and manual transcripts). This network, however, relied on manually generated rather than automatically derived transcripts, limiting its applicability to real-time activity recognition. A keyword-based approach was developed that avoided the need for full transcripts [1] or high-performing ASR in real-time applications. This system built a keyword list for each activity by ranking the frequency of each word and switched the keyword list based on the activity label when spotting keywords. This approach is not practical because it requires knowing the activity label for each utterance to determine the keyword lists. Frequency-based keyword extraction also has limited performance [24]. Frequency-inverse document frequency (TF-IDF) has been used to categorize text documents by weighing each word in text documents based on uniqueness [60, 67, 68]. We used a similar approach in this work, incorporating a sensitivity score that extracts words specific to activity and activity stages from the speech transcripts.

We have observed that activity recognition alone is insufficient for developing effective clinical decision support systems [38, 65]. For example, the utterances ("*Can we get an IV?*") and ("*IV is in place*") refer to the same medical activity (intravenous [IV] catheter placement) but are semantically different. An earlier study [22] successfully recognized speech intentions associated with spoken utterances to determine semantic differences. These speech intentions are too fine-grained with many classes irrelevant to clinical decision support and ignored the relation to the activity type. We build on this prior work by adopting the activity model with four stages of performance, each with distinct speech patterns [30-32]: the "before" stage, such as assessing the need for an activity or requesting an activity; the "during" stage, such as speech related to interactions with patients or reporting activity progress; the "after" stage, such as reports on activity results; and the "other" stage for speech unrelated to activity performance. The utterance ("*Can we get an IV?*") now indicates the "before" stage, while the utterance ("*IV is in place*") indicates the "after" stage. Another utterance ("*I am working on an IV*") indicates the "during" stage. Different activity stages for the same activity can result in different recommendations for decision making. Supplementing activity recognition with the recognition of the activity stage helps avoid semantic misunderstanding, provides additional information about the performed work, improves system recommendations, and brings our system closer to real-time use [65, 69].

To our knowledge, workflow and communication context have not been used for speech-based activity and activity-stage recognition in clinical settings. Context information has been used in natural language processing (NLP) tasks, such as dialog act recognition, where the output is based on the current input and context information. In our setting, a discussion among several team members about an activity may also indicate the status of that activity. A naïve approach would be using a fixed-size context window to collect preceding input to provide better predictions. In the area of text classification, the preceding short texts were used to classify current texts [37, 56]. A fixed context window is not appropriate for scenarios where the occurrence, amount, and length of team communication are unpredictable, and where interruptions and parallel conversations are common. To manage this dynamic context information, a conversational-level model was used to represent the dependencies between different utterances in dialog act recognition [3, 7, 15, 35, 39, 52]. A bi-directional conversational recurrent neural network (RNN) has been used to manage temporal dependencies [35, 39, 52] between utterances. The attention mechanism has also been used to capture the relationships between the utterances in conversation-level model [3, 7, 15]. These models do not easily transfer to our application for two reasons. First, these models were designed for two-people conversations. In contrast, our domain involves team-based communication with a more complex relationship between the utterances (e.g., the conversation might be interleaved or interrupted by another speaker). Second, the sequence-to-sequence models were designed for an offline setting, using the entire conversation log to make a prediction for each utterance. The non-causal nature of these models makes them unfit for online deployment. To overcome these limitations, we designed a causal context module, which independently uses a unidirectional conversational RNN and an attention-based model for interleaved conversation.

The rest of the paper is organized as follows. Section 3 introduces our dataset, including activity and activity-stage labels. Section 4 describes the proposed keyword selection methods and identifies the smallest set of keywords that achieves performance similar to using the entire vocabulary. Section 5 describes our proposed context-aware multimodal network for recognizing activity and activity stages using conversational context from team communication in real time. Section 6 describes a data augmentation method that helps generalize the model's applicability to concurrent activities and activity stages. Section 7 presents the experimental design for offline and online settings, and Section 8 presents the results from offline experiments. Section 9 presents the results from using our model in online settings. Section 10 discusses the results and future work.

3 DATASET DESCRIPTION

Our dataset was recorded during 168 trauma resuscitations in the emergency room at a pediatric hospital between December 2016 and May 2019. The study was approved by the hospital's institutional review board (IRB). The audio data was recorded using two fixed NTG2 Phantom Powered Condenser shotgun microphones with a 16000 Hz sampling rate. These microphones were pointed at two locations around the patient bed, where key members of the team typically stand. All audio recordings were manually transcribed. This preprocessing of audio recordings was time-consuming and required domain knowledge. Censoring the files with patient or provider information involved generating silence in locations of the audio file where this sensitive information was captured. The censoring process was also time-consuming because it required listening to the entire case. After censoring, the recordings were manually segmented into utterances. Although utterance-level segmentation is not available in real-time, it is necessary for model training and accurate label assignment. Utterances were distinct, medically relevant words or sentences spoken by an individual team member and consisted of intelligible speech that could be transcribed. The transcription process followed guidelines to ensure that the transcribed speech could be used for algorithm training. For each utterance, the transcribers marked the start and stop times, the speaker (e.g., emergency medical services (EMS) team member,

Table 1: The statistics of activity-related utterances in our dataset of transcripts from 168 trauma resuscitations.

#	Activity	Code	Utterances
1	No activity performed	None	7,584
2	Head, eye, ear, nose, throat assessment	HEENT	3,117
3	Pre-arrival activities	PR	2,748
4	Extremity examination	E	2,155
5	Medications	MEDS	1,319
6	Back examination	BK	1,464
7	Adjunctive procedures	ADJ	1,327
8	Glasgow coma scale (GCS) evaluation	GCS	1,164
9	Intravenous (IV) catheter placement	IVPLAC	1,155
10	Blood pressure and vital sign measurment	BP	1,004
11	Cervical collar activities	CS	1,077
12	Log roll of patient	LOG	984
13	Circulatory status assessment	CC	905
14	Temperature management (exposure control)	EC	824
15	Airway management	AM	643
16	Summary of reports	SUM	582
17	Airway assessment	AA	571
18	Respiratory assessment (breathing assessment)	BA	504
19	Abdomen examination	A	444
20	Chest palpation	CP	435
21	Cardiopulmonary resuscitation (CPR)-related activities	CPR	434
22	Pelvic examination	PE	432
23	Respiratory support (breathing control)	BC	249

physician examiner, leadership team member, medication nurse, and charge nurse), the transcribed utterance, the associated medical activity, and the activity stage. The utterances were on average 2.3 seconds long.

Prior research using speech for activity recognition in this domain focused only on frequent activities for recognition [1, 24]. Utterances from rare activities are also relevant for resuscitation outcomes and should not be ignored. Supporting even routine decision making requires the recognition of a larger set of activities. For the purposes of our real-time application, a medical expert on our team categorized all clinical tasks into broad groups of related activities (e.g., involving similar procedures or performed on similar body regions) based on standard resuscitation protocol, resulting in 23 activity groups. Each utterance in the transcript was then assigned the corresponding activities. Our dataset shows a “long-tail effect,” with more than nine activities having less than 800 samples (Table 1). This imbalance of utterances across activities presents a challenge for activity recognition. In addition to activity labels, each activity in the utterance was also labeled with its stage [30-32]: before, during, after, or other (Table 2).

Table 2: The statistics of activity-stage-related utterances in our dataset of transcripts from 168 trauma resuscitations.

#	Activity-stage	Utterances
1	Before	5,655
2	During	4,347
3	After	11,307
4	Other	8,857

4 KEYWORD SELECTION AND SPOTTING

Real-time system execution requires automatic speech recognition (ASR) instead of manually generated transcripts. The limited amount of training data and unstructured ambient noise in our domain negatively affected ASR performance. We relied on a small set of keywords for activity recognition, which simplified the problem and removed the need for a real-time high-performing ASR [1]. The goal of keyword selection is to select a small set of words that maintains recognition performance for activities and activity stages. We used two different keyword selection methods—frequency-based and sensitivity-based. We also discuss different keyword spotting methods and introduce our proposed activity-specific training.

4.1 Keyword List

In NLP tasks, standard text preprocessing steps are usually applied to the text before further processing. These steps include removing low-frequency words and stopwords, and text normalization. We followed all text preprocessing steps, except removing stopwords. Stopwords are often removed in the text preprocessing stage because they are not related to the corresponding task. Our evaluation of the activity and activity-stage recognition using the entire vocabulary showed that removal of stopwords pre-defined by NLTK [45] was not needed (Section 8.1.1). Our keyword list was created in two steps. First, we independently created the keyword list for each activity and activity stage. Second, we combined these keyword lists into a common keyword list retaining a single copy of recurring words.

4.2 Frequency-Based Keyword Selection

A frequency-based approach uses the frequency of words associated with each activity to generate a corresponding keyword list. We used a previously described procedure [1] for this purpose we used a cutoff frequency threshold to select the keywords that occurred most frequently with an activity. For example, if 1,000 words are related to an activity and the relative-frequency threshold is 0.1, we can form the keyword list for this activity using the words that appeared more than 100 times. We empirically determined the optimal threshold value to obtain the keyword list. The main limitation of this method is that the keyword list may contain everyday words that appear more frequently than medical terms, leading to omission of medical terms from the keyword list. Although ordinary words may have specific meanings for different activities [1], they are not as useful as medical terms for distinguishing activities and activity stages. This problem could be mitigated by manually filtering ordinary words, but this approach is time-consuming, requires domain knowledge, and may introduce human errors.

4.3 Sensitivity-Based Keyword Selection

Given the limitations of using a frequency-based approach, we considered the "sensitivity measure" to select the words most related to activities, an approach that simultaneously accounts for the frequency and semantic relevance of words. The sensitivity score is adapted from the concept of frequency-inverse document frequency (TF-IDF) defined as:

$$\text{sensitivity}(w|c) = \frac{count(w|C)}{\sum_{w_i} count(w_i|C)} * \frac{count(w|C)}{\sum_{C_i} count(w|C_i)}$$

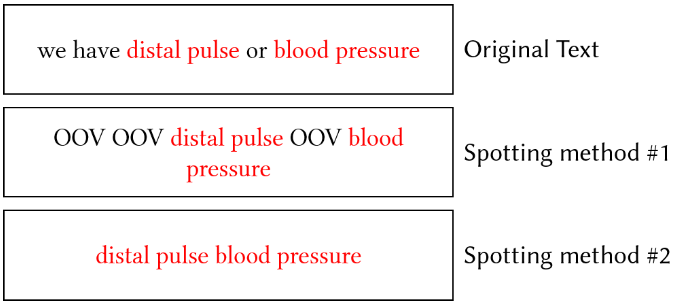
we have distal pulse or blood pressure
Original Text
OOV OOV distal pulse OOV blood pressure
Spotting method #1
distal pulse blood pressure
Spotting method #2

Fig. 2: Examples of different spotting methods. OOV = out-of-vocabulary word.

where w is a given word, C is the given activity or activity-stage category. The first term of the product is the frequency score, which calculates how frequently the word w occurred in the utterances labeled with activity C compared to other words that occurred with C, i.e., how distinguishing the word w is for class C. The frequency score is the same as the frequency-based method (Section 4.2.). The second term is the uniqueness score, which calculates specificity, i.e., how frequently the word w appeared with the activity C compared to all other activities. We multiplied these two terms to make a tradeoff between frequency and uniqueness. Using only the frequency score may lead to overestimating the context relevance of generic words because they would have a high rank in activities where they appeared frequently, but they could also appear with other activities (Section 4.2). Using only the uniqueness score may lead to some words having a high rank even if these words rarely appear. We used this sensitivity measure to rank the words for each activity and empirically set a cutoff sensitivity threshold for keyword selection.

4.4 Keyword Spotting

The keyword spotting stage [1] selects the most frequent word in the keyword list for a given activity and ignores the rest of the list. This model may erroneously assign spotted words that occur in several activities or activity stages. Our approach uses all spotted keywords from the input text because the combination of multiple keywords contains additional cues for activity recognition. We evaluated two keyword spotting approaches (Fig. 2): (1) spotting method #1 - replaces the words that do not appear in the keyword list with an out-of-vocabulary (OOV) symbol, and (2) spotting Method #2 - omits the unspotted words and adds a special symbol “none” if no keywords are spotted in the utterance. Our goal with these different spotting methods was to determine whether the position of the unspotted words affected recognition performance. We used word embedding [46] to represent the keywords instead of the one-hot encoding [1]. We discuss how these methods, including keywords selection and keywords spotting, affected the model performance in our experiments for the context-independent model in Section 8.1.

4.5 Activity-Specific Training

Although using the common keyword list to spot keywords is appropriate, the relationship between keywords and their associated activity and activity stage may be underrepresented when these keywords occur in more than one activity or activity stage. To strengthen the relationship between activity/activity-stage types and associated keywords, we proposed keyword specific training, an approach complementary to keyword spotting methods. We mixed the use of common keyword list with

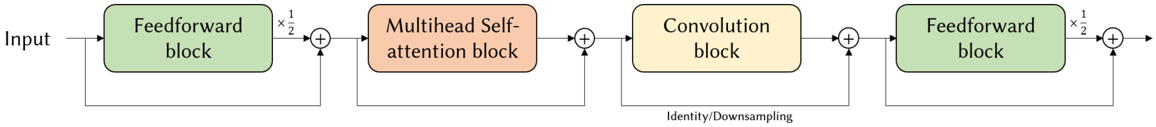

Fig. 3: Conformer block.

the keyword lists specific to the activity for keyword spotting. For example, when we apply the common keyword list to an utterance about medications ("MEDS"), such as "*give another epi right now,*" the following keywords will be spotted: "give," "another," "epi," "right," and "now." Conversely, if a keyword list specific to the activity "MEDS" is used for keyword spotting, only three keywords will be spotted: "give," "another," and "epi." The words "right" and "now" are not activity-specific and should not affect recognition results when omitted. We hypothesized that this training method would strengthen the connections between activity/activity stage and the associated keywords and ensure complete training of the entire keyword list.

5 CONTEXT-AWARE MULTIMODAL MODEL

Our proposed causal context-aware multimodal network was designed to help recognize activities and activity stages based on conversational context from the team communication during trauma resuscitation. This model consists of two parts: (1) an utterance-level model ("context-independent multimodal model") based on a convolutional augmented transformer that uses the spotted keywords and audio features in utterances [25], and (2) a novel context model that learns temporal dependencies in team communication using both unidirectional RNN and attention mechanism.

5.1 Context-Independent Multimodal Model

Our context-independent multimodal model used the spotted keywords and audio features to simultaneously predict the activity and activity stage. The multimodal network used complementary heterogeneous features to provide more accurate predictions [41]. Many studies have investigated the fusion of modal features. Recent methods [23, 61] have focused on aligning different modalities along the time axis, achieving better performance than unaligned methods. Ambient sounds not related to human speech (e.g., alerts) usually come from the instruments or equipment in our domain. For this reason, we chose the late fusion approach that integrates the features processed by each monomodal model.

5.1.1 Monomodal Model

Transformer and its variants have achieved high performance in many domains [17, 18, 25, 47, 63]. The convolution-augmented transformer has outperformed the original transformer because of better modeling of local dependencies [25]. We built our monomodal model based on the conformer block that consists of self-attention, convolution, and feedforward modules (Fig. 3).

A multi-head self-attention with positional encoding is used in the self-attention module. The self-attention calculates an affinity matrix to measure the contribution of inputs to each other. Given an input $X = \{x_1, \ldots, x_N\}$, for every $x_t \in \Re^d, t \in [1, N]$, the attention score between x_t and any other input x_τ is determined as:

$$\alpha_{t\tau} = \frac{\exp\,(\frac{1}{\sqrt{d_k}} x_t^T W_q^T W_k x_\tau)}{\exp\,(\frac{1}{\sqrt{d_k}} \sum_n (x_t^T W_q^T W_k x_n))}$$

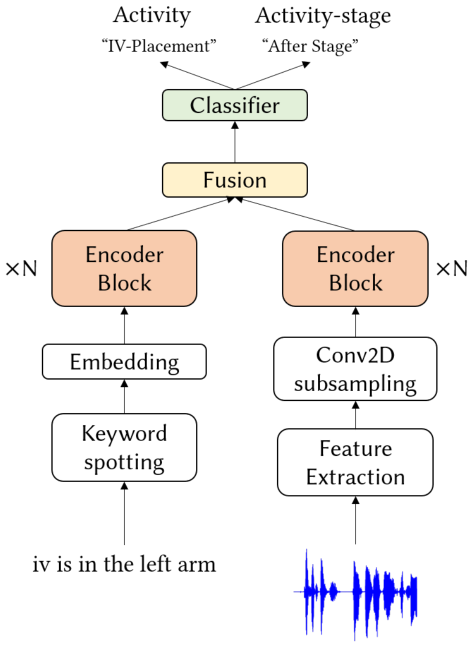

Fig. 4: Context-independent multimodal model.

where $W_q, W_k \in \mathbb{R}^{d \times d_k}$ are the matrices that transform x_t to the *query* space and the *key* space, respectively. The d_k is the dimension of the key and query spaces. $\alpha_{t\tau}$ is the attentive score that measures how x_τ contributes to x_t. For the utterance-level recognition, we did not apply the attention mask in sequence modeling to prevent the disclosure of future information. The final embedding output for x_t is obtained as:

$$z_t = \sum_n \alpha_{tn} W_v x_n$$

where W_v is the transformation that maps the input embeddings to the *value* space. The attention mechanism improves performance by providing a global view of the temporal sequence. The self-attention can be further improved by incorporating a multi-head mechanism that creates multiple attentions to manage complex scenarios. The multi-head self-attention splits the W_q, W_k, W_v along d_k to generate several submatrices, i.e., $W = [w^0; ...; w^h]$, where h is the number of heads. The embedding output for x_t with multi-head attention is obtained as:

$$z_t = W^\circ \left[\sum_n \alpha_{tn}^i W_v^i x_n \right]$$

where $W^\circ \in \mathbb{R}^{d \times d_k}$ is a matrix that aggregates the output from different attention heads and i is the attention head index.

The convolution module uses the separable convolution [13] to enhance the capturing of local relations in the conformer. The input first goes through the step-wise convolution with the gating mechanism [16]. It then follows a depth-wise convolution with the batch norm [29] and swish activation [53]. A stepwise convolution is used for the final output.

The feedforward block enriches the model's representation ability. It is composed of two linear projection layers with an activation function (GELU) in between [26]. A residual connection and layer normalization [6] are used to guarantee a gradient flow for stable training in these three modules.

Finally, the conformer block is built using two feedforward blocks, one self-attention block, and one convolution block (Fig. 3). The monomodal model is a stack of conformer blocks to provide better performance.

5.1.2 Structure of the Context-Independent Multimodal Model

We used two monomodal models for processing keywords and audio signals. The inference speed of the monomodal models prevented real-time application. The input sequence length for different modalities often varied: the number of speech-signal frames was often tenfold the number of words in the corresponding text, leading to slower inference speed for the audio model. We applied downsampling in the audio model to accelerate the inference speed using two approaches: (1) a common subsampling layer was applied before the conformer, and (2) a strided convolution module, which progressively downsamples the features, was used in the conformer block for different modalities [9] instead of the non-strided module [25].

Our context-independent multimodal model used late fusion to integrate the keyword and audio modalities into a shared representation space (Fig. 4). The multimodal model concatenated the output from the keyword and audio monomodal model and passed it through a non-linear transformation. The classifiers were then used to make predictions.

5.2 Causal Context Module

Previous work on speech-based activity recognition used individual utterances for prediction and did not consider the preceding conversational context [1, 22, 24]. Although these models achieved promising results, the performance decreased with an increasing number of activity labels because the individual utterances did not carry sufficient information. Using input from medical experts, we identified three types of verbal communication during trauma resuscitation. The first type is about reporting results from evaluations performed by a provider. Activities associated with these utterances usually can be recognized from the utterance alone, without the conversational context (e.g., "blood pressure is 112 over 85"). The second type includes dialog between team members, like discussion about diagnostic results and the next steps, or synchronizing the teamwork. The third type is a dialog between team members and the patient to assess the patient's condition or obtain information, which requires the patient feedback. The latter two types of verbal communication usually require the context of previous utterances to determine the activity. To help distinguish between the activities, we introduced the context module that captures the dependencies between utterances.

To leverage the context of previous utterances while addressing the limitations discussed in prior work [3, 15, 35, 39, 52], we introduced a causal context module using a unidirectional conversational RNN and attention-based historical conversation interaction to capture continuing and interleaved conversations in a real-time setting.

5.2.1 Unidirectional Conversation RNN

Although a bidirectional conversational RNN showed promising results in dialog act recognition [7, 35, 39, 52], the non-causal property makes it unsuitable for real-time application. For real-time use, we replaced the bidirectional with a unidirectional conversational RNN. We used this unidirectional RNN to capture an ongoing conversation.

5.2.2 Attention-Based Historical Conversation Interaction

Conversational RNN has difficulty managing the dependencies between interleaved conversations that often occur in team communication in our domain. Although the attention mechanism has been used in dialog act recognition, these approaches were either unsuitable for online settings [7, 15], or calculated the attention score using only one previous context vector [3], which could not fully utilize the attention mechanism's ability to model long-range dependencies. Our attention-based historical conversation interaction is a stand-alone module, which uses the encoder-decoder attention. It is defined as:

$$y_t = \mathit{Enc\text{-}DecAttn}([u_t], [u_1 \dots u_{t-1}])$$

where u_t represents the output from the context-independent multimodal model at time *t*. The attention-based historical conversation interaction stored the preceding utterance-level representations as historical dialogue, which allowed it to retrieve long-past historical information. The current utterance-level representation was used as the query for the encoder-decoder mechanism. The attention-based historical conversation interaction managed the interleaved-style conversations while preserving the causal property for the real-time model.

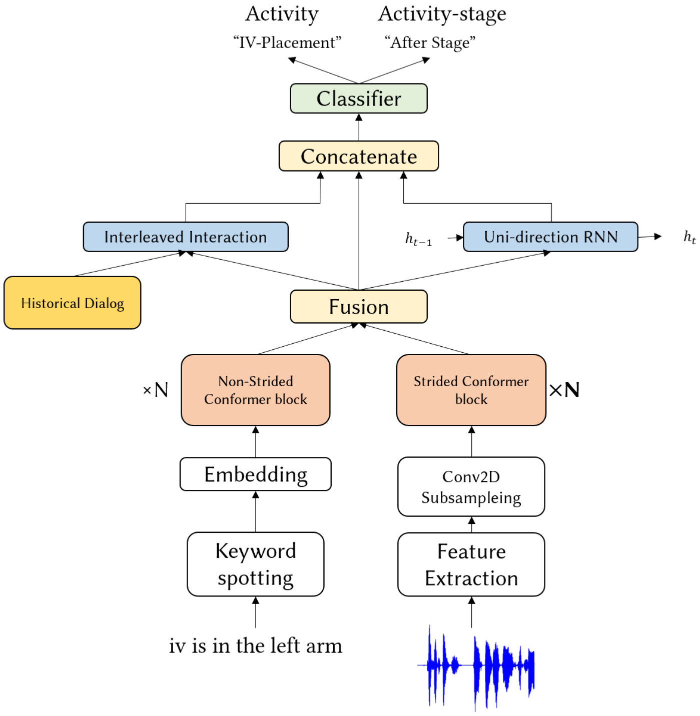

Fig. 5: Context-aware multimodal model.

5.2.3 *Structure of the Context-Aware Multimodal Model*

The two context modules formed our proposed context module (Fig. 5). These two modules were used independently to capture different kinds of dependencies. The context module was applied on top of the context-independent multimodal model. The concatenation of the output of the context-independent model and the output of the context module were used to make the final predictions.

6 DATA AUGMENTATION

Our data were transcribed and segmented into utterances for accurate label assignment. The transcribers transcribed most speech into utterances about individual activities. For a small subset, segmenting chunks of text based on individual activities was not possible, leading to a small number of utterances related to multiple activities and activity stages. The utterance-level segmentation is not available in a real-time setting. We used a fixed-length sliding time window of speech signal for recognizing the activities and activity stages. The time window could not capture enough information if the window length was too short because utterances were fragmented. If the time window was too long, it would cause a delay in real-time recognition, which is not expected. For use in a real-time setting, we choose the window length of five seconds as a tradeoff between the fragmentation problem and delays. Because this window length (5 seconds) was longer than the average length of utterance-level segmentation (2.3 seconds), utterances within the window were more likely to refer to multiple activities and activity stages. To enhance the model's ability to recognize concurrent activities and their stages, we introduced two data augmentation methods. We first simulated context-independent utterances with multiple activities and stages. Second, we simulated team communication during trauma resuscitation. Our data augmentation method is described below. We justify the feasibility of data augmentation for both offline experiments and real-time application in Sections 8 and 9.

6.1 Simulating Context-independent Utterances

The goal of simulating context-independent utterances was to strengthen the model's ability to simultaneously recognize multiple activities. We composed new utterances by randomly selecting the utterances from our training dataset. Some utterances can only be recognized within a given context. Selecting these utterances may confuse the training of the model. To avoid this issue, we identified a set of context-independent utterances and assumed these could be recognized without contextual information because they preserved enough information for recognition. We identified context-independent utterances based

```
Algorithm 1 Simulate Team Communications
Input: Actual prearrival- and postarrival-related conversations from the training corpus
Output: Simulated trauma conversation
function SimulateTraumaConversation($Pre\_arrival, Post\_arrival, max\_len$)
    $case$ = [Random_Select($Pre\_arrival$)]
    while $len(case) < max\_len$ do:
        $conversation$ = Random_Select($Post\_arrival$)
        $conversation$ = Insert_or_Append($conversation, single\_utter$)
        $case.append(conversation)$
    end while
    return $case$
end function
```

Fig. 6: Workflow for simulating team communication during a trauma case

on activity/activity stage specific training (Section 4.5). If activity-specific keywords were spotted in an utterance then we classify this utterance as context-independent. Finally, we simulated new utterances by randomly selecting and combining these context-independent utterances.

6.2 Simulating Team Communication During Actual Trauma Resuscitations

In addition to simulating context-independent utterances, we simulated team communication for data augmentation (Fig. 6). We identified two subsets of conversations from available transcripts in several steps. First, we collected all utterances that formed a conversation about each activity in a given resuscitation. We started with the activity related to the first utterance of a transcript and traversed the subsequent utterances until an utterance about a different activity was encountered. The previous sentence was considered as the termination of the first conversation. The resulting set of adjacent utterances and the associated audio clips formed an activity-related conversation. We then used the activity in the current utterance as the start of the next conversation and continued in this way until the end of the transcript. Finally, we divided the conversations into the subsets related to patient pre-arrival and post-arrival phases. We simulated team communication during resuscitation cases using these steps (Fig. 6):

1. We selected several pre-arrival conversations as the beginning of a simulated team communication.
2. Repeat either (a) or (b)
    a) We appended a context-independent utterance (either simulated as in Section 6.1 or an original) to a randomly-selected conversation that occurred after patient arrival.
    b) We inserted a context-independent utterance into a randomly-selected conversation that occurred after patient arrival to represent an interleaved conversation.
3. We repeated Step 2 until the simulated team communication reached the maximum specified number of utterances.

7 EXPERIMENTAL SETUP

We trained and evaluated both the context-independent and context-aware models (described in Section 5). For offline experiments, we first compared different keyword generation and spotting methods using the context-independent model. After identifying the optimal keyword list and the spotting method, we evaluated the proposed context-aware model and performed an ablation study. We then ran online experiments to evaluate model performance in an online setting. We also performed experiments to show the efficiency of data augmentation with offline and online settings.

7.1 Dataset Preparation for Offline Experiments

We trained and evaluated the proposed models using 168 trauma resuscitation transcripts with 28,468 utterances (Table 1). Because of the small dataset, we applied five-fold cross-validation based on the unit of individual transcripts. We separated a fraction of transcripts (10%) from the training set to serve as the validation set. Each data sample (an utterance) contained a unique identifier indicating the sequential order of the utterance from the start of the resuscitation, a normalized human-transcribed utterance (text cleaned from the punctuations, Arabic numerals converted to text, and text normalization and contraction expansion), an audio clip, and the corresponding activity and activity-stage labels. Keyword lists were generated based on the entire dataset using different generation methods (Sections 4.2 & 4.3). We used a window size of 25 ms with a shift of 10 ms for commonly used 80-dimensional filterbank feature extraction from audio. We evaluated the efficiency of our proposed data augmentation with different numbers of simulated utterances obtained from the training set, including 5,000, 10,000, 15,000, and 20,000 sample size.

7.2 Label Alignment for System Evaluation in Online Experiments

We used the context-aware model trained with data collected offline to evaluate its performance in real time. For real-time application, we used an input based on a five-second sliding time window instead of utterance-level segmentation. The system-generated label predictions were stored for post-event validation. After the event, we used a manually generated transcript with utterance-level segmentation to assign the activity and activity-stage labels to each time-window of speech. For each time-window, we checked if it overlapped with time intervals generated from the start-end timestamps of the manually segmented utterances. We omitted utterances that partially overlapped with the given time window (shorter than one-third of the original utterance interval and shorter than one half of the time window). We then merged the activity and activity-stage labels for the remaining overlapped utterances. We used these aligned labels to evaluate our system in an online setting.

7.3 Model Implementation and the Training Configuration

We implemented our models using Pytorch [51] 1.8.0 with one Nvidia GTX 2080 GPU and CUDA version 10.1. Adam [34] was used as an optimization algorithm, setting the initial learning rate as 1e-3. Dropout [57] was set as 0.15 in all experiments to avoid overfitting. No other regularization or procedures were applied. We used a batch size of 256 utterances in all experiments. The random seed was fixed for reproducibility. We used binary cross entropy as the loss function to update our model for multilabel classification. To address the imbalance of the dataset (Section 3), we calculated weights based on statistical analysis of the dataset and applied the weights to the binary cross entropy to rescale the penalty for different activities and activity stages. The run time for the five-fold validation was about two days. We used F_1-score as the evaluation metric. During training, we halved the learning rate if the weighted F_1-score among all activities in the validation set did not increase in three consecutive epochs. Early stop criteria were triggered if the F_1-score could not be improved in 15 epochs.

7.4 Baseline Models

To provide a strong baseline for a fair comparison of our proposed approach, we designed several baseline models using existing techniques. Bidirectional Encoder Representations from Transformers (BERT) [17] leverages large amounts of unannotated data by using self-supervised pretraining. BERT has shown promising performance by fine-tuning the pretrained model in various natural-language tasks. In audio domains, this technique has been known as self-supervised representation learning. We used the HuBERT [66] to extract audio feature, with the pretrained model showing generalizability to various audio tasks. We designed three models based on these techniques:

- Model #1: A monomodal model that used the BERT base model [17] (contains around 110 million trainable parameters) pretrained with BooksCorpus [70] and English Wikipedia, and its pretrained tokenizer. We added two classifiers that used the representation generated by BERT for activity and stage predictions.
- Model #2: Model #1 modified to replace BERT with a pretrained Clinical BERT base model [4], trained with a publicly available clinical-domain corpus that addresses the domain-shifting. We also used the Clinical BERT base and its pretrained tokenizer for fine-tuning.

- Model #3: A multimodal model that concatenates the audio representation extracted by HuBERT [28] with the text representation extracted by either BERT or Clinical BERT. We selected either a BERT or Clinical BERT text representation based on which one performed better in model #1 and model #2. We then fused different modalities and used two classifiers to predict activity and activity stage. We did not use audio representation alone because it did not contain sufficient cues to distinguish between activities and activity stages. Audio could be used as a supplementary modality to improve the recognition performance.

Table 3: Performance comparison between using the original vocabulary and filtered vocabulary. The numbers in the parentheses indicate the vocabulary size (in words). Weighted average F_1-score is reported.

Selection Method	Activity	Activity-Stage
Original Vocabulary (4,717)	65.7±1.4%	75.4±1.2%
Filtered Vocabulary (4,591)	64.8±0.9%	71.8±1.6%

8 RESULTS FROM THE OFFLINE EXPERIMENTS AND DISCUSSION

We first selected the optimal keyword list and model configuration for context-independent model (Section 8.1). We then evaluated our proposed context-aware model (Section 8.2). The data augmentation experiments are presented in Section 8.3. Finally, Section 8.5 presents our analysis.

8.1 Experimental Configuration and Results for the Context-Independent Model

We performed a series of experiments for the context-independent model. First, we showed that regular stopwords are needed in our study domain. Second, we evaluated different keyword selection methods and identified the best keyword list and spotting method. Finally, we performed an ablation study that compared the model configuration and our proposed activity-specific training. We used the best-performing context-independent model for father experiments.

8.1.1 Evaluating the Impact of Removing Regular Stopwords from Vocabulary

We performed two experiments to evaluate the performance of the context-independent model with and without stopwords in the keyword list. In the first experiment, we used the entire vocabulary. In the second experiment, we used a filtered vocabulary where we removed regular stopwords that were introduced by NLTK [45]. We used the same configuration for the context-independent model in both experiments, setting up a 3-layer context-independent model with a word embedding size of 256 and the number of heads in multi-heads attention set at 8. We used the spotting method #1 (described in Section 4.4) and repeated the five-fold validation. We then independently compared the weighted F_1-score among the different categories for activity and activity stage (Table 3). The recognition performance for activity and activity stage decreased when we removed the regular stopwords (the activity had a 0.9% performance decline and activity stage had a 3.6% performance decline). This result showed that regular stopwords improved activity and activity-stage recognition, suggesting they should not be removed in advance. In the remaining experiments, we did not remove these regular stopwords from the keyword lists.

Table 4: Frequency and sensitivity thresholds and vocabulary sizes. When a threshold was 0, we used the entire vocabulary.

Selection Method	0.01	0.005	0.0025	0.0015	0.001	0.00075	0.0005	0.00025	0.0001	0.0
Frequency-based	140	237	371	500	661	768	999	1651	2881	4717
Sensitivity-based	139	248	340	436	512	729	1100	2240	2948	4717

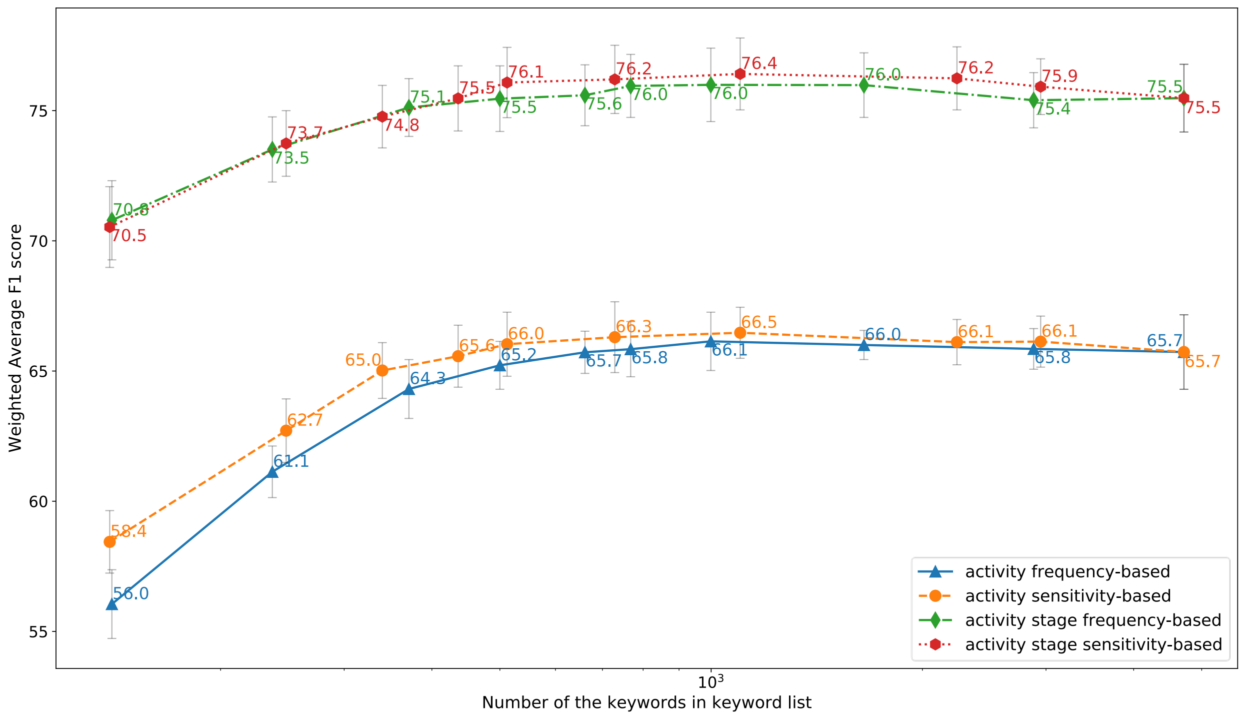

Fig. 7: Activity and Activity-stage recognition weighted F_1-score for 5-fold validation using different keyword lists.

8.1.2 *Selecting the Methods for Keyword List Generation*

Our goal was to minimize the vocabulary size (keyword list) needed for activity and activity-stage recognition while preserving high system performance (Section 4). In these experiments, we set a series of thresholds that resulted in different vocabulary sizes (Table 4). We did not enforce the same vocabulary size for frequency and sensitivity-based selection methods because we thresholded statistical parameters for these different selection methods (relative frequency versus sensitivity score). Because we were studying the effects of keywords selection, we used the same model configuration (described in Section 8.1.1). We repeated the five-fold validation for each keyword list.

For activity recognition results (Fig. 7), we observed that the sensitivity-based keyword selection method performed better than the frequency-based method for vocabularies with less than 1,000 words. As the vocabulary size increased, the performance differences between different selection and spotting methods decreased. The performance improvement was small when the vocabulary exceeded 2,000 words because most domain-specific keywords were already included when smaller-size keyword selections were used.

For activity-stage recognition results (Fig. 7), we observed that frequency-based method had nearly identical performance compared to sensitivity-based method when the vocabulary size was smaller than 500. The sensitivity-based method performed slightly better when the vocabulary size increased (greater than 500).

Table 5: The results for ablation study of Context-independent model. Two-step training first uses activity-specific keyword lists, followed by the common list.

Model Configuration*	Spotting Method	Modality	Activity/Activity-stage specific spotting	Activity	Activity-stage
Post-norm, 256, 3	#1	Text+Audio		66.3±1.0%	76.1±1.3%
Post-norm, 256, **6**	#1	Text+Audio		66.0±1.4%	76.1±1.9%
Post-norm, **384**, 3	#1	Text+Audio		66.2±1.3%	76.0±1.6%
Pre-norm, 256, 3	#1	Text+Audio		66.1±1.1%	75.9±1.2%
Post-norm, 256, 3	#1	Text+Audio	•	66.7±0.9%	76.3±1.2%
Post-norm, 256, 3	#2	Text+Audio		65.9±1.1%	75.3±1.2%
Post-norm, 256, 6	#2	Text+Audio	•	66.4±1.1%	75.5±1.3%
Post-norm, 256, 3	#1	Text	•	64.9±1.2%	75.2±1.8%

* The model configuration in the table represents pre- or post-norm style, the dimension size for word embedding, and the number of layers used in the model.

We also observed that the performance for activity and activity-stage recognition using an appropriate-size keyword list had better performance compared to using the entire vocabulary, suggesting that keyword-based methods are more robust in the recognition tasks with limited-size data. We chose the keyword list based on the sensitivity method with a vocabulary size of 729 words (“sensitive-729”) to make a tradeoff between recognition performance and the keyword list size. We used this keyword list in the remaining experiments.

8.1.3 Ablation Study for the Context-Independent Model

After identifying the optimal keyword list, we performed an ablation study to find the best setting for the context-independent model. Our study included the effects of various hyperparameters on the model configuration, different keyword-spotting methods (Section 4.4), the proposed activity-specific training (Section 4.5), and the effect of different modalities. Changes in the model configuration included modifying the number of layers used in the encoder (3 vs. 6), the dimension of text embedding (256 vs. 384), and positional embedding. We also selected between the pre-norm (as in [25]) or post-norm style (our adaption) of the encoder, and between a single modality (text) or combined modalities (text+tudio) as input. The audio modality alone was not considered because aforementioned limitations. These different model configurations had only a small impact on the recognition performance (Table 5). The two spotting methods achieved comparable performance for activity recognition. Spotting method #1 outperformed spotting method #2 on activity-stage recognition because the unspotted keywords from method #1 provided additional position information in the text, helping distinguish the difference in activity stages. Our proposed activity/activity-stage specific training strengthened the relations between activity, activity stage, and associated keyword lists. The ablation study of different modalities showed that the multimodal models outperformed the single-modality model (text) in both activity and activity-stage recognition (Table 5, last row).

Based on the results from the ablation study, we chose the original setting (post-norm, 256, 3) with spotting method #1 and used activity-specific training in the remaining experiments.

8.2 Experimental Configuration and Results for the Context-Aware Model

8.2.1 Comparing Context-Independent and Context-Aware Models

We next performed the experiments with the proposed context-aware multimodal model. The context-aware model outperformed the context-independent model by a large margin for activity recognition (66.7% versus 71.5%) and had a slight improvement in performance for activity-stage recognition (76.3% versus 77.6%). To understand how the context-aware model made these improvements, we independently analyzed the performance for each activity and activity stage and compared it with the context-independent model (Table 6, columns 1 & 4). We highlighted in bold face the activity (HEENT, PR, E, BK, GCS, IVPLAC, LOG, SUM, A) and activity stage (During) when major improvements were made using the context-aware model. The recognition results for these activities and activity stages are affected by the conversational context, showing that conversational context is needed for recognizing specific types of activities and activity stages. The improvement for the remaining activities and activity stages was lower, implying a smaller effect of conversational context.

8.2.2 *Ablation Study for the Context-Aware Model*

We next performed an ablation study to assess how each of the three context modules contributed to the recognition of activities and activity stages. We constructed three models (Table 6): a model that only contained conversational-RNN (RNN), a model that only contained attention-based historical conversation interaction (Attention), and our proposed model that combined the first two (context-aware model). The results for each class are shown independently (Table 6, columns 3, 4, and 5). By comparing the results for each activity and activity-stage, we observed that: (1) the RNN model performed better than the Attention model for these activities, including HEENT, E, BK, CS, GCS. The conversations for these activities are usually continuous and short-

Table 6: F_1-scores (in percentages) for activities and activity stages using the context-aware model with three designs (last three columns). Boldfaced numbers highlight significantly improved performance compared to the context-independent model (first column).

Model Configuration	Context-independent	RNN	Attention	Context-aware
Weigted F_1-scores for All Activities	66.7±0.9	70.9±1.0	71.1±1.0	**71.5±1.2**
No activity performed (None)	67.8±2.5	70.0±2.0	69.5±2.5	70.0±2.6
Head, eye, ear, nose, throat assessment (HEENT)	**71.1±3.6**	76.1±2.9	75.3±2.7	**76.2±3.1**
Pre-arrival activities (PR)	**56.0±3.8**	65.4±3.1	67.5±4.1	**68.6±3.5**
Extremity examination (E)	**66.1±5.4**	76.1±3.0	75.1±3.7	**75.9±2.8**
Medications (MEDS)	71.6±4.7	72.7±5.7	74.9±4.5	74.5±6.7
Back examination (BK)	**51.9±2.3**	71.3±4.1	70.1±4.2	**71.4±3.6**
Adjunctive procedures (ADJ)	68.4±2.6	70.7±2.9	69.7±4.3	70.5±3.1
Glasgow coma scale (GCS) evaluation (GCS)	**66.8±2.3**	72.3±2.9	71.5±2.7	**72.6±2.5**
Intravenous (IV) catheter placement (IVPLAC)	**69.6±4.3**	74.5±3.4	74.4±2.7	**74.6±2.6**
Blood pressure and vital sign measurment (BP)	75.7±3.6	76.8±3.2	76.8±3.4	77.8±2.8
Cervical collar activities (CS)	60.3±4.7	63.6±5.7	62.9±3.7	63.5±3.7
Log roll of patient (LOG)	**69.8±5.4**	76.3±4.3	76.5±4.0	**76.4±4.0**
Circulatory status assessment (CC)	75.5±4.3	76.0±3.3	76.5±3.1	76.1±3.5
Temperature management (exposure control) (EC)	74.7±1.5	77.3±2.6	77.3±0.7	77.4±1.0
Airway management (AM)	53.3±5.4	57.4±7.7	57.9±6.9	57.9±7.5
Summary of reports (SUM)	**44.2±4.7**	46.6±6.4	49.0±8.3	**50.5±5.5**
Airway assessment (AA)	77.7±3.6	77.7±4.0	78.2±3.6	77.7±4.3
Respiratory assessment (breathing assessment) (BA)	67.1±4.1	68.9±3.5	66.3±3.0	70.0±4.1
Abdomen examination (A)	**66.6±5.9**	71.8±7.4	73.1±3.5	**73.6±5.0**
Chest palpation (CP)	71.9±5.3	72.2±6.1	74.4±2.9	73.9±5.4
Cardiopulmonary resuscitation-related activities (CPR)	58.2±2.2	61.0±4.3	61.3±2.5	62.2±1.9
Pelvic examination (PE)	67.1±3.2	71.4±3.8	71.2±2.6	71.8±3.0
Respiratory support (breathing control) (BC)	48.7±6.8	49.9±6.1	50.0±6.4	50.1±6.6
Weigted F_1-scores for All Activity Stages	76.3±1.2	77.6±1.2	77.5±1.0	77.6±1.0
Before	68.9±2.3	69.2±2.3	69.4±2.2	69.2±2.1
During	**77.5±2.5**	81.7±3.0	81.4±2.7	**81.8±2.9**
After	80.6±2.2	81.7±2.1	81.6±2.0	81.7±2.0
Other	73.9±2.0	75.6±1.5	75.5±1.5	75.8±1.6

Table 7: Performance comparison between without-data-augmentation and data-augmentation. Weighted Average F_1-scores for activity/activity-stage recognition are reported.

	Context-independent	**Context-aware**
W/O Data Augmentation	66.7±0.9/76.3±1.2%	71.5±1.2/77.6±1.0%
Data Augmentation (5,000)	67.4±0.9/76.4±1.1%	72.3±1.2/77.9±0.9%
Data Augmentation (10,000)	68.0±1.1/76.4±1.2%	72.9±0.9/77.9±1.0%
Data Augmentation (15,000)	68.5±1.2/76.4±1.2%	73.2±0.8/78.1±1.1%
Data Augmentation (20,000)	68.6±1.3/76.3±1.2%	73.2±0.7/78.2±1.2%

duration because providers take the initiative to perform activities and report results. RNN is better at handling these continuous local conversational dependencies; (2) the Attention model performed better than the RNN model for some activities, including PR, SUM, A, and CP. These activities either had longer-duration (PR, SUM), or required feedback from patients (A, CP). The long durations or possible delayed responses were the main causes of interleaved conversations, when other activities were more likely to overlap and break the continuity of the communication. The Attention model was better at handling these discontinuous dependencies in conversation; (3) both RNN and Attention achieved comparable performance in activity-stage recognition; and (4) the context-aware model performed better than the RNN and Attention models. These results were observed because our model combined the advantages of both RNN and Attention models, allowing it to handle complex team communication as the observations (1) and (2) were not always true in real scenarios. For example, the GCS activity is usually performed by a single provider and is a "continuous" conversation. Sometimes GCS is assigned to several providers and is associated with a "discontinuous" conversation, interleaved with other activities. In another example related to activity A, a delayed response from the patient may cause an interleaved conversation. The providers did not need to wait for the response when visible injuries were observed or the patient lost consciousness. Based on these results, we used the context-aware model in the remaining experiments.

8.3 Data Augmentation and its Effect for Offline Settings

Data augmentation improved the performance of the context-aware model and partially the performance of the context-independent model (Table 7). We empirically analyzed the effectiveness of the proposed data augmentation methods, gradually increasing the number of simulated samples from 5,000 to 20,000. Models trained with data augmentation performed better than those trained without data augmentation. This performance improvement first increased as the number of simulated data grew and then converged to performance comparable to when augmented with 15,000 or 20,000 samples. For the rest of our analysis, we used the data augmentation with 15,000 simulated samples because it most improved the performance with the fewest simulated samples. Data augmentation improved the performance for both activity (1.7%) and activity-stage (1.5%) recognition in the context-aware model. For the context-independent model, data augmentation improved activity recognition (1.8%), but contributed less to activity-stage recognition (0.1%). Data augmentation was more effective for the context-aware model because it improved the recognition of the "during" activity stage. These findings suggest that our proposed data augmentation improved the performance of both models in the offline setting.

Table 8: Performance comparison to baselines. Weighted Average F_1-scores for activity/activity-stage recognition are reported.

		W/O Data Augmentation	Data Augmentation (15,000)
Baselines	BERT (text)	65.0±1.0/76.0±1.6%	–
	ClinicalBERT (text)	65.5±1.1/76.1±1.5%	–
	ClinicalBERT+HuBERT (text+audio)	67.3±0.7/76.6±1.3%	68.7±0.9/76.9±1.2%
Our models	Context-independent (text)	64.9±1.2/75.2±1.8%	–
	Context-independent (text+audio)	66.7±0.9/76.3±1.2%	68.5±1.2/76.4±1.2%
	Context-aware (text+audio)	71.5±1.2/77.6±1.0%	73.2±0.8/78.1±1.1%

8.4 Comparisons to Baseline Models

We compared our proposed approaches to three baseline models. We first evaluated the single-modality baseline Models #1 and #2 (BERT and Clinical BERT). The performance of activity stages improved in both BERT and Clinical BERT compared to the context-independent (text) model (Table 8). We propose that pretraining with a large amount of unannotated data helped the model distinguish the semantic differences in activity stages. BERT achieved activity-recognition performance comparable to the context-independent (text) model. Clinical BERT showed an improvement in activity-recognition performance, proving more helpful to activity recognition in clinical domains. For the multimodal baseline Model #3, we combined text representation from Clinical BERT with audio presentation from HuBERT [28]. We observed that this combination improved recognition performance, showing how different modalities influence performance in our domain. The improvement with ClinicalBERT+HuBERT over ClinicalBERT was about the same as the improvement of our context-independent (text+audio) over context-independent (text) (Table 8). We suspect that this comparable performance improvement was caused by instruments sounds that helped activity recognition more than human speech (and HuBERT was pretrained with clean human speech). Applying data augmentation to ClinicalBERT+HuBERT further improved the performance and showed that our proposed data augmentation improved performance in a limited-data scenario (last column in Table 8). Data augmentation showed less performance improvement when the model (ClinicalBERT+HuBERT) was pretrained with a large unannotated dataset compared to the model trained from scratch (context-independent). Although the ClinicalBERT+HuBERT baseline outperformed the context-independent model, it had two limitations. First, the ClinicalBERT+HuBERT baseline has more parameters than the context-independent model, leading to high computational costs. Second, the ClinicalBERT+HuBERT baseline ignores the conversational context, which led to a better performance of the context-aware model (last row in Table 8). Despite these limitations, the baseline models showed promise for improving context-independent model by self-supervised learning using a large unannotated in-domain corpus. This approach will be evaluated in our future work.

8.5 Analysis and Discussion

We analyzed the context-aware model performance for each activity and activity stage and conducted an error analysis for the low-performing activities (CS, AM, SUM, CPR, BC). The main error-type was misclassifying the activity group as "none," particularly for the activities AM and CPR (Table 6). Activities BC and BA were sometimes confused, and SUM was often misclassified with its finer-level activities, such as A, AA, BA, and E. For the remaining activities, the context-aware model performed well (more than 70% of F_1-scores) (Table 6). In activity-stage recognition, the main confusion occurred between the

Table 9: Comparing our context-aware system performance with prior work [1] on their set of activities. F_1-score is reported.

Activity / Model	Extremity examination (E)	Back examination (BK)	GCS evaluation (GCS)	Head, eye, ear, nose, throat assessment (HEENT)	Circulatory status assessment (CC)
Previous System [1]	63.2%	66.5%	56.8%	59.9%	56.7%
Our System	75.9±2.8%	71.4±3.6%	72.6±2.5%	75.3±3.3%	75.7±3.5%

"before" and "other" stages. Some "other" stage utterances were related to requesting items or information that were unrelated to medical activities (e.g., "What time is it?"). When comparing the model performance for a subset of activities used in previous research [1], our system significantly outperformed the previous work (Table 9).

We also jointly evaluated the activity and activity-stage recognition considering four scenarios (Table 10): (1) both activity and activity stage were correctly recognized (TT), (2) the activity was correctly recognized while activity stage was incorrectly recognized (TF), (3) the activity was incorrectly recognized while activity stage was correctly recognized (FT), and (4) both activity and activity stage were incorrectly recognized (FF). In most cases, activity and activity stage were simultaneously recognized correctly (TT), but in a subset of cases only one was recognized correctly (TF or FT). To ensure accurate decision support recommendations, activity and activity stage need to be correctly recognized. If the clinical decision support system is used to recommend future activities in the workflow, an incorrect activity-stage recognition could lead to an incorrect recommendation. To achieve high reliability in the actual application, our speech-based activity recognition system should be combined with other modalities, such as computer vision for detecting activity stage [41]. Tracking the workflow and predicting future activities could also be achieved based on learning from the past cases [65].

9 RESULTS FROM THE ONLINE EXPERIMENTS AND DISCUSSION

We also evaluated the context-aware model in online experiments. We built and trained our ASR model using one training set from a five-fold validation (Section 9.2). We performed two types of online experiments and compared their results to the offline setting. First, we ran the experiments using ASR output in place of manual transcripts with the utterance-level segmentation. We then replaced the utterance-level segmentation with time-window segmentation and validated the model in a real-time setting.

9.1 Training of the ASR

Recent deep learning-based ASR [8, 11, 25, 54, 63] has provided reliable performance. To replace offline manual transcripts, we built a customized end-to-end ASR model using ESPNet toolkit [64]. We used the pretrained WavLM [12], which was trained using self-supervised learning with large-scale training dataset as the front-end in the acoustic model. We set the acoustic model

Table 10: Confusion matrix of the correctness (true/false) for activity and activity-stage recognition.

Activity-Stage / Activity	T (Activity-stage correctly recognized)	F (Activity-stage incorrectly recognized)
T (Activity correctly recognized)	52.7±1.6%	11.1±1.0%
F (Activity incorrectly recognized)	19.6±1.0%	15.8±0.9%

as a 6-layer encoder, followed by a 3-layer decoder based on the conformer architecture [25]. A 6-layer transformer was used as the language model [63]. We used one set of a five-fold validation for training and testing in the ASR experiment. We used raw data augmentation with the Librispeech corpus [49] and spectrum augmentation [50] during training. This customized ASR achieved a 46.7% of word error rate (WER) in the testing set. The main contributor to the high WER was the ambient noise, reverberation, and low-volume speech. These factors and the limited size of the training dataset negatively affected ASR performance.

9.2 Replacing Manual Transcripts with the ASR Output

Although the performance of the ASR was inadequate, we evaluated substituting the manual transcripts with the ASR output as the input for our *keyword-based* activity and activity-stage recognition. We compared the performance of activity and activity-stage recognition for two context-aware models. The first model was trained on the entire vocabulary and the second on a selected keyword list ("sensitive-729" from Section 8.1.2). We then segmented the audio based on the utterance-level segmentation from the manual transcripts and replaced the manual transcript with the ASR output. When using the manual transcripts, our keyword-based model performed marginally better than the model based on the entire vocabulary (Table 11, row 1). Data augmentation further improved performance (Table 11, third column). When using the ASR output to replace the manual transcripts (Table 11, row 2), the system performance declined by about 10%. The keyword-based model performed better than the model based on the entire vocabulary, most likely because the keyword-based model was less affected by ASR performance. These findings suggest that the keyword-based model is more robust and that data augmentation improved performance. The proposed models (Table 11, columns 2-4) with the context module all outperformed the baseline model (ClinicalBERT+HuBERT) with data augmentation (Table 11, last column). This result showed that capturing the conversational context is critical in activity and activity-stage recognition during complex workflows such as trauma resuscitation. The pretrained tokenizer in [4, 17] was more negatively affected by a poorly performing ASR than our word-level model. This may be because the pretrained tokenizer would encode the out-of-vocabulary words that were wrongly generated from poorly performing ASR, negatively impacting recognition results.

9.3 Real-Time Activity and Activity-Stage Recognition

Although our context-aware model achieved moderate performance after replacing manual transcripts with the ASR output, the utterance-level segmentation is not available in real time. To evaluate our model in a real-time setting, we used a time-window segmentation. We aligned the utterance-level labels with each time window (Section 7.2) and measured the system performance in a real-time setting (Table 11, last row). The performance of each model further declined when relying on the time-window segmentation and compared to aligned labels. The main cause of this decline may be utterance fragmentation that occurred during the label alignment. For example, an utterance may span two time-windows, but the keywords associated with the activity may only occur in the second window. Because of the causal property for the real-time model, the information for the first window is inadequate for the model to recognize the corresponding activity. The performance further improved with data augmentation. This finding suggests that concurrent activities and activity stages occurred more frequently in fixed-length time windows. Our model achieved reasonable performance for a real-time setting such as dynamic and complex medical workflow (Table 11, shaded in grey). This baseline model (ClinicalBERT+HuBERT) performed worst because it ignored the conversational context and was more affected by a poorly performing ASR when using a tokenizer.

Table 11: Performance comparison between manual transcripts and the ASR output. Weighted Average F_1-scores for activity and activity-stage recognition are reported. The number of words used in each case is shown in the first row.

	Entire vocabulary (4717)	**Keyword-based (729)**	**Keyword-based (729) with data augmentation**	**ClinicalBERT+HuBERT with data augmentation**
Utterance-level segmented & manual transcripts	69.3/75.1	69.6/76.8	**72.0/77.4**	68.1/74.8
Utterance-level segmented & ASR	58.7/61.2	60.0/62.7	**60.9/65.2**	54.1/59.3
Time window sliding & ASR	53.9/58.1	55.4/59.9	**57.4/61.6**	50.9/55.2

10 CONCLUSION AND FUTURE WORK

We developed a system for recognizing 23 types of medical activities and 4 activity stages. We introduced a dynamic keyword spotting and selection method that relies on a small set of keywords. Using multiple data sources, such as audio sound, keywords, and the context of previous sentences, we developed a context-aware multimodal architecture to recognize activities and corresponding activity stages. This proposed system addresses the following requirements for an online model: (1) alleviating the need for a high-quality ASR; (2) capturing the conversational context during a complex workflow; (3) recognizing both activities and activity stages, a required feature for developing a clinical decision support system. We have built this system specific to trauma resuscitation scenario, but it can be generalized to the scenarios where activity recognition is also based on verbal communication and had common routine activities, such as surgery and ICU. We introduced a data augmentation method to improve performance and evaluated the system using manually transcribed speech (offline experiments) and real-time speech (online experiments). In offline experiments, the results for 23 activity types and 4 activity stages showed an average F_1-score of 73.2 % and 78.1% for activity and activity stage recognition, respectively. In online experiments, our model performance declined by about 10% in both activity and activity-stage recognition when using utterance-level segmentation with the ASR output and declined around 15% when using time-window sliding with the ASR output. Although models were trained with utterance-level data and performed on time-window segments, online experiments showed the feasibility of using speech to detect activity and activity stage in complex medical workflows.

The performance of our model may still be inadequate for some activities in a real-time decision-support system due to the poorly performing ASR and a limited training dataset. Our future work will focus on complementing speech-based activity recognition with computer vision-based methods. Some activities for which our speech-based system showed lower accuracy (CPR, BC) have clear visual cues and could be reliably recognized using computer vision. Our future work will consider four approaches to improve the performance of our current system: (1) improving the training objective for imbalanced datasets; (2) designing a more realistic data augmentation method; (3) pretraining the model with a large amount of in-domain unannotated data using self-supervised learning; (4) improving the ASR performance by semi- and self-supervised learning to supplement a limited-size human-annotated dataset.

ACKNOWLEDGMENTS

This research has been funded under NIH/NLM grant 2R01LM011834-05 and NSF grant IIS-1763827.

REFERENCES

[1] Jalal Abdulbaqi, Yue Gu, Zhichao Xu, Chenyang Gao, Ivan Marsic, and Randall S Burd. 2020. *Speech-Based Activity Recognition for Trauma Resuscitation*, in *IEEE International Conference on Healthcare Informatics (ICHI).* IEEE, Oldenburg, Germany, p. 1-8. https://doi.org/10.1109/ICHI48887.2020.9374372.

[2] Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. *Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions*, in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.* **5**(1), p. 1-24. https://doi.org/doi.org/10.1145/3448090.

[3] Ali Ahmadvand, Jason Ingyu Choi, and Eugene Agichtein. 2019. *Contextual dialogue act classification for open-domain conversational agents*, in *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, p. 1273-1276. https://doi.org/10.1145/3331184.3331375.

[4] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. *Publicly available clinical BERT embeddings*, in *2nd Clinical Natural Language Processing Workshop.* Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 72--78. https://doi.org/10.18653/v1/W19-1909.

[5] Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. 2010. *Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey*, in *23th International conference on architecture of computing systems 2010.* VDE, p. 1-10.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. *Layer normalization*, in *arXiv preprint arXiv:1607.06450.* https://doi.org/10.48550/arXiv.1607.06450.

[7] Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter. 2018. *Conversational analysis using utterance-level attention-based bidirectional recurrent neural networks.* INTERSPEECH. https://doi.org/10.21437/Interspeech.2018-2527.

[8] Herve A Bourlard and Nelson Morgan, *Connectionist speech recognition: a hybrid approach.* Vol. 247. 2012: Springer Science & Business Media.

[9] Maxime Burchi and Valentin Vielzeuf. 2021. *Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition*, *arXiv preprint arXiv:2109.01163.* 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), p. 8--15. https://doi.org/10.1109/ASRU51503.2021.9687874.

[10] Ishani Chakraborty, Ahmed Elgammal, and Randall S 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, p. 1-8. https://doi.org/10.1109/FG.2013.6553758.

[11] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. *Listen, attend and spell*, in *arXiv preprint arXiv:1508.01211.* 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP.2016.7472621.

[12] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, and Xiong Xiao. 2021. *Wavlm: Large-scale self-supervised pre-training for full stack speech processing*, in *IEEE Journal of Selected Topics in Signal Processing.* https://doi.org/10.1109/JSTSP.2022.3188113.

[13] François Chollet. 2017. *Xception: Deep learning with depthwise separable convolutions*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 1251-1258. https://doi.org/10.1109/CVPR.2017.195.

[14] Yohan Chon, Nicholas D Lane, Fan Li, Hojung Cha, and Feng Zhao. 2012. *Automatically characterizing places with opportunistic crowdsensing using smartphones*, in *Proceedings of the 2012 ACM conference on ubiquitous computing*, p. 481-490. https://doi.org/doi.org/10.1145/2370216.2370288.

[15] Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloe Clavel. 2020. *Guiding attention in sequence-to-sequence models for dialogue act prediction*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, p. 7594-7601. https://doi.org/10.1609/aaai.v34i05.6259.

[16] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. *Language modeling with gated convolutional networks*, in *International conference on machine learning.* PMLR, p. 933-941. https://doi.org/10.48550/arXiv.1612.08083.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding.* Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 4171--4186. https://doi.org/10.18653/v1/N19-1423.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. 2020. *An image is worth 16x16 words: Transformers for image recognition at scale.* International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2010.11929.

[19] Mark Fitzgerald, Peter Cameron, Colin Mackenzie, Nathan Farrow, Pamela Scicluna, Robert Gocentas, Adam Bystrzycki, Geraldine Lee, Nick Andrianopoulos, and Linas Dziukas. 2011. *Trauma resuscitation errors and computer-assisted decision support*, in *Archives of Surgery.* **146**(2), p. 218-225. https://doi.org/10.1001/archsurg.2010.333.

[20] Mark Fitzgerald, Rob Gocentas, Linas Dziukas, Peter Cameron, Colin Mackenzie, and Nathan Farrow. 2006. *Using video audit to improve trauma resuscitation—time for a new approach*, in *Canadian journal of surgery.* **49**(3), p. 208.

[21] 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*, in *Annals of surgery.* **244**(3), p. 371. https://doi.org/10.1097/01.sla.0000234655.83517.56.

[22] Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. *Speech intention classification with multimodal deep learning*, in *Canadian conference on artificial intelligence.* Springer, p. 260-271. https://doi.org/10.1007/978-3-319-57351-9_30.

[23] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. *Multimodal affective analysis using hierarchical attention strategy with word-level alignment*, in *Proceedings of the conference. Association for Computational Linguistics. Meeting.* NIH Public Access, p. 2225. https://doi.org/10.18653/v1/P18-1207.

[24] Yue Gu, Ruiyu Zhang, Xinwei Zhao, Shuhong Chen, Jalal Abdulbaqi, Ivan Marsic, Megan Cheng, and Randall S Burd. 2019. *Multimodal attention network for trauma activity recognition from spoken language and environmental sound*, *2019 IEEE International Conference on Healthcare Informatics (ICHI).* IEEE, p. 1-6. https://doi.org/10.1109/ICHI.2019.8904713.

[25] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, and Yonghui Wu. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition.* INTERSPEECH. https://doi.org/10.21437/Interspeech.2020-3015.

[26] Dan Hendrycks and Kevin Gimpel. 2016. *Gaussian error linear units (gelus)*, in *arXiv preprint arXiv:1606.08415.* https://doi.org/10.48550/arXiv.1606.08415.

[27] John B Holcomb, Russell D Dumire, John W Crommett, Connie E Stamateris, Matthew A Fagert, Jim A Cleveland, Gina R Dorlac, Warren C Dorlac, James P Bonar, and Kenji Hira. 2002. *Evaluation of trauma team performance using an advanced human patient simulator for resuscitation training*, in *Journal of Trauma and Acute Care Surgery.* **52**(6), p. 1078-1086. https://doi.org/10.1097/00005373-200206000-00009.

[28] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* **29**, p. 3451-3460. https://doi.org/10.1109/TASLP.2021.3122291.

[29] Sergey Ioffe and Christian Szegedy. 2015. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, in *International conference on machine learning.* PMLR, p. 448-456. https://doi.org/10.48550/arXiv.1502.03167.

[30] Swathi Jagannath, Neha Kamireddi, Katherine Ann Zellner, Randall S Burd, Ivan Marsic, and Aleksandra Sarcevic. 2022. *A Speech-Based Model for Tracking the Progression of Activities in Extreme Action Teamwork*, in *Proceedings of the ACM on Human-Computer Interaction.* **6**(CSCW1), p. 1-26. https://doi.org/10.1145/3512920.

[31] Swathi Jagannath, Aleksandra Sarcevic, Neha Kamireddi, and Ivan Marsic. 2019. *Assessing the Feasibility of Speech-Based Activity Recognition in Dynamic Medical Settings*, in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 1-6. https://doi.org/10.1145/3290607.3312983.

[32] 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*, p. 88-97. https://doi.org/10.1145/3240925.3240941.

[33] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. 2019. *Human activity recognition: A survey*, in *Procedia Computer Science.* **155**, p. 698-703. https://doi.org/10.1016/j.procs.2019.08.100.

[34] Diederik P Kingma and Jimmy Ba. 2014. *Adam: A method for stochastic optimization*, in *arXiv preprint arXiv:1412.6980.* https://doi.org/doi.org/10.48550/arXiv.1412.6980.

[35] Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. *Dialogue act sequence labeling using hierarchical encoder with crf*, in *Proceedings of the aaai conference on artificial intelligence.* https://doi.org/10.1609/aaai.v32i1.11701.

[36] Nicholas D Lane, Petko Georgiev, and Lorena Qendro. 2015. *Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning*, in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, p. 283-294. https://doi.org/doi.org/10.1145/2750858.2804262.

[37] Ji Young Lee and Franck Dernoncourt. 2016. *Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks*, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 515-520. https://doi.org/10.18653/v1/N16-1062.

[38] Keyi Li, Sen Yang, Travis M Sullivan, Randall S Burd, and Ivan Marsic. 2022. *Exploring Runtime Decision Support for Trauma Resuscitation*, in *arXiv preprint arXiv:2207.02922.* https://doi.org/10.48550/arXiv.2207.02922.

[39] Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. *A Dual-Attention Hierarchical Recurrent Neural Network for Dialogue Act Classification*, in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, p. 383-392. https://doi.org/10.18653/v1/K19-1036.

[40] Xinyu Li, Dongyang Yao, Xuechao Pan, Jonathan Johannaman, JaeWon Yang, Rachel Webman, Aleksandra Sarcevic, Ivan Marsic, and Randall S Burd. 2016. *Activity recognition for medical teamwork based on passive RFID*, in *2016 IEEE international conference on RFID (RFID).* IEEE, p. 1-9. https://doi.org/10.1109/RFID.2016.7488002.

[41] Xinyu Li, Yanyi Zhang, Mengzhu Li, Shuhong Chen, Farneth R Austin, Ivan Marsic, and Randall S Burd. 2016. *Online process phase detection using multimodal deep learning*, in *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON).* IEEE, p. 1-7. https://doi.org/10.1109/UEMCON.2016.7777912.

[42] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall S Burd. 2016. *Deep learning for rfid-based activity recognition*, in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, p. 164-175. https://doi.org/10.1145/2994551.2994569.

[43] Xinyu Li, Yanyi Zhang, Jianyu Zhang, Yueyang Chen, Huangcan Li, Ivan Marsic, and Randall S Burd. 2017. *Region-based activity recognition using conditional GAN*, in *Proceedings of the 25th ACM international conference on Multimedia*, p. 1059-1067. https://doi.org/10.1145/3123266.3123365.

[44] Dawei Liang and Edison Thomaz. 2019. *Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos*, in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.* **3**(1), p. 1-18. https://doi.org/doi.org/10.1145/3314404.

[45] Edward Loper and Steven Bird. 2002. *NLTK: The Natural Language Toolkit*, in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, p. 63-70.

[46] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient estimation of word representations in vector space*, in *arXiv preprint arXiv:1301.3781.* https://doi.org/10.48550/arXiv.1301.3781.

[47] Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer. 2019. *Transformers with convolutional context for ASR.* https://doi.org/10.48550/arXiv.1904.11660.

[48] Ed Oakley, Sergio Stocker, Georg Staubli, and Simon Young. 2006. *Using video recording to identify management errors in pediatric trauma resuscitation*, in *Pediatrics.* **117**(3), p. 658-664. https://doi.org/10.1542/peds.2004-1803.

[49] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. *Librispeech: an asr corpus based on public domain audio books*, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, p. 5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964.

[50] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. *Specaugment: A simple data augmentation method for automatic speech recognition.* INTERSPEECH. https://doi.org/10.21437/Interspeech.2019-2680.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. 2019. *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in neural information processing systems.* **32**, p. 8026-8037. https://doi.org/10.48550/arXiv.1912.01703.

[52] Vipul Raheja and Joel Tetreault. 2019. *Dialogue act classification with context-aware self-attention.* Association for Computational Linguistics, Minneapolis, Minnesota, USA, p. 3727--3733. https://doi.org/10.18653/v1/N19-1373.

[53] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. *Searching for activation functions*, in *arXiv preprint arXiv:1710.05941.* https://doi.org/10.48550/arXiv.1710.05941.

[54] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. 2018. *Light gated recurrent units for speech recognition*, in *IEEE Transactions on Emerging Topics in Computational Intelligence.* **2**(2), p. 92-102. https://doi.org/10.1109/TETCI.2017.2762739.

[55] Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. 2020. *Automatic operating room surgical activity recognition for robot-assisted surgery*, in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, p. 385-395. https://doi.org/10.1007/978-3-030-59716-0_37.

[56] Xingyi Song, Johann Petrak, and Angus Roberts. 2018. *A deep neural network sentence level classification method with context information.* Association for Computational Linguistics, p. 900--904. https://doi.org/10.18653/v1/D18-1107.

[57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. *Dropout: a simple way to prevent neural networks from overfitting*, in *The journal of machine learning research.* **15**(1), p. 1929-1958.

[58] Hang Su, Wen Qi, Chenguang Yang, Jiehao Li, Xuanyi Zhou, Giancarlo Ferrigno, and Elena De Momi. 2020. *Human Activity Recognition Enhanced Robot-Assisted Minimally Invasive Surgery*, in *International Conference on Robotics in Alpe-Adria Danube Region.* Springer, p. 121-129. https://doi.org/10.1007/978-3-030-48989-2_14.

[59] ATLS Subcommittee and International ATLS Working Group. 2013. *Advanced trauma life support (ATLS®): the ninth edition*, in *The journal of trauma and acute care surgery.* **74**(5), p. 1363-1366. https://doi.org/10.1097/TA.0b013e31828b82f5.

[60] Bruno Trstenjak, Sasa Mikac, and Dzenana Donko. 2014. *KNN with TF-IDF based framework for text categorization*, in *Procedia Engineering.* **69**, p. 1356-1364. https://doi.org/10.1016/j.proeng.2014.03.129.

[61] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. *Multimodal transformer for unaligned multimodal language sequences*, in *Proceedings of the conference. Association for Computational Linguistics. Meeting.* NIH Public Access, p. 6558. https://doi.org/10.18653/v1/P19-1656.

[62] Donald D Vernon, Ronald A Furnival, Kristine W Hansen, Edma M Diller, Robert G Bolte, Dale G Johnson, and J Michael Dean. 1999. *Effect of a pediatric trauma response team on emergency department treatment time and mortality of pediatric trauma victims*, in *Pediatrics.* **103**(1), p. 20-24. https://doi.org/10.1542/peds.103.1.20.

[63] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, and Frank Zhang. 2020. *Transformer-based acoustic modeling for hybrid speech recognition*, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, p. 6874-6878. https://doi.org/10.1109/ICASSP40776.2020.9054345.

[64] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, and Nanxin Chen. 2018. *Espnet: End-to-end speech processing toolkit*, in *arXiv preprint arXiv:1804.00015.* INTERSPEECH. https://doi.org/10.21437/Interspeech.2018-1456.

[65] Sen Yang, Xin Dong, Leilei Sun, Yichen Zhou, Richard A Farneth, Hui Xiong, Randall S Burd, and Ivan Marsic. 2017. *A data-driven process recommender framework*, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, p. 2111-2120. https://doi.org/10.1145/3097983.3098174.

[66] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, and Guan-Ting Lin. 2021. *Superb: Speech processing universal performance benchmark*, in *arXiv preprint arXiv:2105.01051.* INTERSPEECH. https://doi.org/10.21437/Interspeech.2021-1775.

[67] Zhang Yun-tao, Gong Ling, and Wang Yong-cheng. 2005. *An improved TF-IDF approach for text classification*, in *Journal of Zhejiang University-Science A.* **6**(1), p. 49-55. https://doi.org/10.1007/BF02842477.

[68] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. *A comparative study of TF* IDF, LSI and multi-words for text classification*, in *Expert systems with applications.* **38**(3), p. 2758-2765. https://doi.org/10.1016/j.eswa.2010.08.066.

[69] Yanyi Zhang, Ivan Marsic, and Randall S Burd. 2021. *Real-time medical phase recognition using long-term video understanding and progress gate method*, in *Medical Image Analysis.* **74**, p. 102224. https://doi.org/10.1016/j.media.2021.102224.

[70] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*, in *Proceedings of the IEEE international conference on computer vision*, p. 19-27. https://doi.org/10.1109/ICCV.2015.11.