Speech-Based Activity Recognition for Trauma Resuscitation

2020 IEEE International Conference on Healthcare Informatics (ICHI) | 978-1-7281-5382-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHI48887.2020.9374372

Jalal Abdulbaqi
Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey
Piscataway, NJ, USA
jalal.nazar@rutgers.edu

Yue Gu
Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey
Piscataway, NJ, USA
yue.guapp@rutgers.edu

Zhichao Xu
Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey
Piscataway, NJ, USA
zhichao.xu@rutgers.edu

Chenyang Gao
Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey
Piscataway, NJ, USA
chenyang.gao@rutgers.edu

Ivan Marsic
Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey
Piscataway, NJ, USA
marsic@rutgers.edu

Randall S. Burd
Trauma and Burn Surgery
Children's National Medical Center
Washington, DC, USA
rburd@childrensnational.org

***Abstract*— We present a speech-based approach to recognize team activities in the context of trauma resuscitation. We first analyzed the audio recordings of trauma resuscitations in terms of activity frequency, noise-level, and activity-related keyword frequency to determine the dataset characteristics. We next evaluated different audio-preprocessing parameters (spectral feature types and audio channels) to find the optimal configuration. We then introduced a novel neural network to recognize the trauma activities using a modified VGG network that extracts features from the audio input. The output of the modified VGG network is combined with the output of a network that takes keyword text as input, and the combination is used to generate activity labels. We compared our system with several baselines and performed a detailed analysis of the performance results for specific activities. Our results show that our proposed architecture that uses Mel-spectrum spectral coefficients features with a stereo channel and activity-specific frequent keywords achieve the highest accuracy and average F1-score.**

Keywords—activity recognition, keyword, audio classification, speech processing, trauma resuscitation

I. Introduction

Activity recognition of a dynamic medical process such as trauma resuscitation is challenging because of fast and concurrent work as well as a noisy environment. There are several current research approaches that rely on video, RFID, and signals from medical devices to identify the medical activity type and its stages [1]–[3]. However, to our knowledge, there have been no approaches that rely on the speech from verbal communication of the team. Video and RFID data cannot provide information to recognize certain activities. For instance, in the trauma resuscitation, Glasgow coma score calculation (GCS) and airway assessment (AA) activities rely on visual examination or talking to the patient and can be recognized only based on verbal communication. We asked three medical experts to rate different modalities (speech, video and RFID-tagged object) as the best source for recognizing different ongoing resuscitation activities. In Table I, we averaged their ratings for four activities for which speech was rated the highest as a prediction source had they been asked to do activity recognition. In addition, a study [4] found that medical experts can predict resuscitation activities with 87% accuracy using only the verbal communication transcripts. Furthermore, previous studies showed that fusing the speech with the video, RFID or transcripts increases activity recognition accuracy [5], [6].

We present a speech-based activity recognition design for dynamic medical teamwork and empirical evaluation. Our approach is based on using one representative keyword from the input utterance to the activity recognition network, in addition to the audio stream. This keyword belongs to the most frequent words list that has been calculated for each activity type. In addition, to determine the challenges related to system design and dataset limitations, we determined the dataset characteristics related to the activities (e.g. activity frequency, noise-level and word frequency for each activity). Then, we analyzed different audio preprocessing parameters, such as feature types and input channels to find the best input feature setup. Using these findings, we designed an audio classification network based on the VGG model [7]. We evaluated our audio network and

TABLE I. Activities for Which Three Medical Experts Rated Highest Speech as The Modality for Activity Recognition

Activity	Audio (%)	Video (%)	RFID tag
GCS Calculation	80	7.5	Non
Airway Assessment	80	20	Non
Medications	80	20	Partial
CPR	65	45	Partial

compared its performance with several state-of-the-art classification networks using the trauma resuscitation dataset. Finally, we evaluated our keyword-based network design using different settings for the network layers. We found that a keyword-based approach to activity recognition performed better than relying on manually-generated transcripts. The results show that our new keyword-based design increased the accuracy and the average F1-score by 3.6% and 0.184 respectively compared to our audio network alone. Our contributions are:

- An analysis of trauma resuscitation dataset characteristics to determine the constraints related to speech-based activity recognition.
- Audio preprocessing analysis to find the optimal parameters for designing the network.
- Design of an audio classification network and comparison of its performance to the state-of-the-art classification models using a trauma resuscitation audio dataset.
- A new keyword-based neural network for activity recognition that combines the audio stream and the most frequent words from the input transcript.

The rest of the paper is organized as follows: Section II analyzes the dataset attributes. We describe the audio preprocessing in Section III and describe the model design in detail in Section IV. The experimental setup and the results are presented in Section V. We review the related work in Section VI. We conclude and propose future work in Section VII.

II. Related Work

In recent years, activity recognition for medical purposes has been growing quickly. Most of the current research relies on the sensors and visual modalities such as the passive RFID and the videos, and there are few works based on audio and verbal information.

RFID-based activity recognition considered an object-use detection problem. Early work compared different machine learning approaches as a binary classifier to predict the medical object motion that related to certain activities [8]. A different strategy to place the RFID tags showed an improvement in the activity recognition accuracy [9], [10]. Recently, employing a convolutional neural network (CNN) as a multi-class classifier outperformed the previous approaches [2]. Although, RFID has advantages such as being small, cheap and battery-free, its accuracy and scalability is limited by the radio noise and the limited number of activities that use taggable objects.

Visual-based activity recognition exploits the visual data from RGB or depth camera to map the medical team movement and actions into activities. Early research used a single camera video recording with the Markov Logic Network model to predict the activities [1]. Recently, deep learning has been applied to visual-based activity recognition. The convolutional neural network has been applied for video classification using time-stacked frames with a slow fusion network to process the short-range temporal association of activities [11]. To address the short-range temporal limitation, a long short term memory network (LSTM) has been integrated with the VGG network with a region-based technique to generate an activity mask [12]. Despite the decent progress in utilizing virtual data, it has several limitations. The RBG camera raises the issue of patient privacy and, as with the RFID, not all activities can be predicted by visual tracking the medical team movement and actions.

Text-based activity recognition employed the transcript of the verbal communication between the medical team to predict the activity type. Recent research applied a multi-head attention architecture [13] to predict a speech-reliant activity from the transcripts and the environmental sound [6]. The drawback in this approach is that obtaining the text requires additional automatic speech recognition (ASR).

The audio modality was used as an auxiliary to other modalities in works [5], [6]. These papers analyzed the audio ability to improve the accuracy of the activity recognition. In [5], the authors built a multimodal system to recognize concurrent activities by using multiple data modalities: depth camera video, RFID sensors and audio recordings. Each modality processed and the features extracted by a separate convolutional neural network (CNN), and then all of them fused using Long Short-Term Memory (LSTM) network to the final decision layer. They did not provide quantitative analysis to distinguish the difference between each modality performance. In [6], the authors created a multimodal transformer network to process the transcribed spoken language and the environmental sound to predict the trauma activities. The quantitative analysis showed the average accuracy 36.4 when using only audio, and the accuracy increased to 71.8 when using both modalities.

III. Dataset Collection and Characteristics

The dataset was collected during 86 trauma resuscitations in the emergency room at a pediatric teaching hospital in the U.S. Mid-Atlantic region between December 2016 and May 2017. We obtained approvals from the hospital's Institutional Review Board (IRB) before the study. All data generated during the study were kept confidential and secure in accordance with IRB policies and Health Insurance Portability and Accountability Act (HIPAA). The audio data was recorded using two fixed NTG2 Phantom Powered Condenser shotgun microphones. These microphones pointed in two locations where the key members of a trauma team normally stand. The recordings were manually transcribed and each sentence was assigned the activity label by trauma experts. In this section, we present an analysis of the following three characteristics of the dataset that can affect the activity recognition outcome: activity frequency, noise level and words frequency for each activity.

The fine-grained activities have been grouped into 30 high-level categories. Different categories occurred with different frequencies, which is the total number of utterances that include a given activity category for the 86 resuscitations cases (Table II). As seen, the activities are not distributed uniformly over the dataset utterances. Some activities occurred very frequently, while others were rare. There are several reasons for this variation. First, the length of conversation between the medical team is different for each activity. Some activities require several inquiries and reports, while other activities may have a single sentence to report the patient's status. Second, each patient required different evaluation and management activities based

TABLE II. RESUSITATION ACTIVITIES WITH MOST UTTERANCES

#	Activity	Code	Utterances
1	Extremity	E	836
2	Back	BK	701
3	GCS Calculation	GCS	610
4	Face	F	514
5	Circulation Control	CC	407
6	Log Roll	LOG	389
7	C-Spine	CS	380
8	Medications	MEDS	358
9	Pulse Check	PC	289
10	Blood Pressure	BP	256
11	Ear Assessment	EAR	246
12	Eye Assessment	EY	246
13	Exposure Control	EC	220
14	Abdomen Assessment	A	208
15	Breathing Assessment	BA	206
16	Airway Assessment	AA	197
17	Head	H	175
18	Exposure Assessment	EA	174
19	CPR	CPR	160
20	Chest Palpation	CP	150
21	Breathing Control	BC	137
22	Pelvis Assessment	PE	122
23	LEADS	LEADS	116
24	Endotracheal Tube Endorsement	ET	109
25	Neck Assessment	NECK	96
26	Intubation	I	50
27	Genital Assessment	G	44
28	NGT	NGT	30
29	Bolus	B	18
30	Relieve Obstruction	RO	13
	Total activity-labeled utterances		7457

on the patient injury, demographics and medical context. Finally, as mentioned above, the activity categories are a high-level groups that sometimes include several low-level activities (Table III). Hence, when an activity group (e.g. Extremity Assessment) has several low-level activities, this tends to correspond to increased verbal communication of the medical team. As a result of this non-uniform activity distribution, it is hard to train a neural network model for the activities that had associated least-frequent utterances, even for activities that cannot be recognized from other modalities (e.g. Airway Assessment), because of insufficient data to train and evaluate the model. Therefore, we chose the top five activities that had associated highest-frequency utterances in Table II for the purpose of our experiments: Extremity, Back, GCS Calculation, Face, and Circulation Control. All other utterances that do not belong to these activity categories are assigned to the "OTHER" category.

TABLE III. FOUR HIGH-LEVEL ACTIVITIES AND THEIR RELATED LOW-LEVEL ACTIVITES

High-Level Activity	Low-level Activity
GCS Calculation	Verbalized
	Motor Assess
	Verbal Assess
	Eye Assess
Extremity Assessment	Right Upper
	Left Upper
	Right Lower
	Left Lower
Medications	Medications
Airway Assessment	Airway Assessment
CPR	Chest comp
	Shock
	Defib pads
	ID

The second important dataset parameter that can influence the recognition performance is the ambient noise. The resuscitation environment presents several challenges to speech-based activity recognition. Concurrent speakers ("cocktail party" problem), rapid speech and ambient noise adversely affect the speech quality and reduce activity recognition accuracy. To estimate the clarity of the medical team speech, we performed a subjective evaluation of the trauma resuscitation dataset. In this evaluation, we categorized the noisiness of audio recordings into three levels based on the human ability to understand the reports of patient vital signs and examination results. Three medical experts worked on this assessment listening to the 86 resuscitation cases. Each case had been labeled with one of the three noise categories (low, medium and high) and the average is shown in Table IV. As seen, about 65% of cases were labeled as low-noise, while about 19% and 16% were labeled as a medium- and high-level, respectively. Thus, about 35% in our dataset are either unintelligible or it is hard to understand what the medical team said during the resuscitation, which is challenging for the neural network performance. To study the effect of the ambient noise on the recognition of our selected activities, we calculated the number of noisy cases for each chosen activity (Fig.1). Fig. 1 shows the fraction of the resuscitation cases by their noise level for each activity. As seen, the noise is distributed almost uniformly among the activities in our experiment. Therefore, it is not expected to affect some activities more than others in terms of prediction accuracy.

The keywords about patient medical status are the most important information of the team verbal communication in the trauma resuscitation, which sometimes indicates the activity explicitly (e.g. GCS in Fig. 2). To find the priority of the keywords with respect to the related activities, we first filtered most of the stop words. Then, we calculated the most-frequent words for each activity (Fig. 2). As seen, most of the shown keywords either have a direct relationship with the activity (e.g. "spine" for BACK) or have indirect meaning such as the body position (e.g. right or left for extremity). Also, we can see that there are several words that have no relationship with the activities, but they frequently occur as a part of the inquiry response (e.g. "okay") or just part of a repeated sentence (e.g. "get"). However, our intuition is that as long as these words occur frequently during certain activity then these words are valuable for the neural network to predict the correct activity. Hence, our hypothesis is that these keywords can be combined with the audio stream and fed into the neural network to increase the activity recognition accuracy. Extracting these keywords can be done automatically using a word-spotting model. We believe that extracting such keywords is easier and more efficient than

TABLE IV. A SUBJECTIVE EVALUATION OF NOISE FOR ALL 86 RESUSCITATION CASES BY THREE RATERS

Noise Level	Number of Cases
High	14
Medium	16
Low	56
Total	86

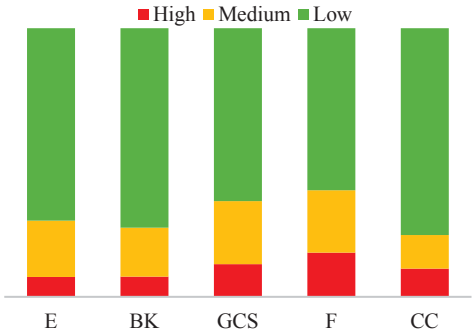

Fig. 1. Cases noise-level distribution for each activity.

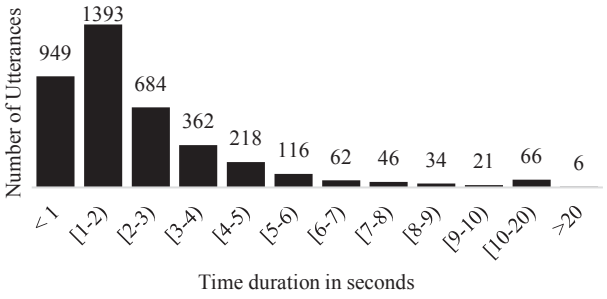

Fig. 3. Utterance-level audio length distribution.

recognizing the whole utterances using an ASR system. In this work, we evaluated the concept of combining given text keywords and the audio stream to improve activity recognition performance. Although evaluating a word-spotting model is not part of this work, we will consider that in our future research.

IV. Data Preprocessing and Configuration

Our main input data is the utterance-level audio stream. In addition, we considered using one keyword from the most-frequent words list as an additional input. The keyword input is encoded as a one-hot one-dimensional vector, and the audio stream is converted into a spectrogram. Spectrogram representation reduces the dimension of the data and provides better information representation [14]. This section describes the data preprocessing and an analysis of two parameters variation effect on the activity recognition outcomes: feature type and input channels.

The keyword feature represented as a one-hot vector of size 78 to represent the total 60 words list (10 keywords per activity). The one-hot vector size had been incremented by 0.3 to reduce the one-hot hash function collision probability. The audio recordings were sampled at 16MHz. We used 40 filter banks for the short-time Fourier transform with a 2048 window, 25% overlap and Hann window type. The audio stream utterances had different time lengths (Fig. 3). The average utterance time duration was 2.42 seconds with a standard deviation of 2.28. Our neural network required a fixed input length, which can be implemented in several ways. First, we could choose a small input size that most of the utterences have such as 1-2 sec or 2-3 sec, but this would reduce the total number of samples. Second, we could specify a fixed length such as the average value and then truncate all the longer utterances, but our experiments showed that the lost information would significantly reduce the performance. Therefore, we resized all the utterance lengths to be 20 seconds by zero-padding the beginning and end of each utterance. The final feature map

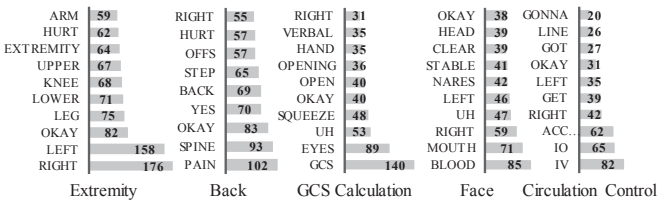

Fig. 2. The most frequent unique words for each activity.

length was 600. Following the work [6], we segmented the input feature map into 10 frame sub-maps to avoid processing distant audio frames. The input sample shape for every single channel was (60, 40, 10).

We tried two types of audio spectrogram feature: Mel-frequency cepstral coefficients (MFCC) and Mel-frequency spectral coefficients (MFSC). MFCC feature extraction has been successfully applied in speech recognition [15] and audio classification [16]. However, MFCC includes the discrete cosine transform (DCT), which can compromise the locality, especially for the convolutional neural network (CNN) [17]. Therefore, several audio classification works used MFSC instead [18]. Furthermore, we analyzed the effect of adding the dynamic features: the first and second temporal derivatives (delta and the delta-delta coefficients, respectively). Adding the dynamic features can increase the accuracy and the robustness of speech recognition [19]. Table V shows the evaluation results for both feature types and their derivatives. The results show that MFSC features dominante over the MFCC with and without their derivatives. The reason is the locality issue introduced by DCT of MFCC transform mentioned above. Also, we noticed that adding the derivatives to the MFCC feature type slightly increased the accuracy, while adding the derivatives for the MFSC degraded the accuracy. Therefore, we concluded that the static MFSC is the best feature type for our dataset and architecture, and we used it in the next experiments.

As mentioned in Section II, our audio data were recorded using two microphones. As a consequence, each audio recording included two channels. We combined the two channels in five different configurations (Table VI). In the first two configurations, we used only one of the channels. In the third setup, we doubled the dataset by feeding both channels as a distinct samples. In the last two setups, we used both channels together either by summing them up and averaging into a single channel, or by feeding them as a two-channels. Table VII shows the accuracy and the average F1-score for each configuration. The accuracy of the last two setups, when the two channels are combined, is higher than the first three setups when the input is

TABLE V. The Accuracy and Average F1-Score for Different Features Types

Feature type		Accuracy	Average F1-Score
MFCC	Static	26.0	0.162
	Dynamic (Δ, ΔΔ)	27.7	0.200
MFSC	Static	**30.8**	**0.231**
	Dynamic (Δ, ΔΔ)	30.0	0.210

TABLE VI. INPUT CHANNEL CONFIGRATION

Input Configuration	Number of Samples	Input Dimension
CH1 only	3557	(60, 40, 10)
CH2 only	3557	(60, 40, 10)
Unite CH1 with CH2	3557×**2**	(60, 40, 10)
(CH1 + CH2)/2	3557	(60, 40, 10)
Combine CH1 & CH2	3557	(60, 40, **20**)

one channel only. The reason is that the labels were transcribed based on both channels, so when one of input channels is omitted, some utterances may have wrong labels and consequently the neural network failed to predict the activity on the evaluation dataset. Combining the two-channels achieved higher accuracy. However, the average of the two channels had slightly lower accuracy than including both channels. Thus, in the next evaluation experiments, we considered the configuration that used the static MFSC feature type and feeding the network with both channels.

V. MODEL ARCHITECTURES

We considered the speech-based activity recognition as a multi-class classification problem. This section first presents a modified VGG [7] network for the audio branch, which we used to evaluate the configurations described in Section III. Then, we introduced a new architecture that fuses the output of the proposed audio network with the keyword network to predict the activities.

A. The Audio Network

Previous neural network architectures designed for image processing have been adjusted successfully to work on audio processing [18], [20] such as VGG [7], ResNet [21] and DenseNet [22]. The VGG network shows a better performance compared to other architectures for audio classification applications [18], [23]. Because deeper CNN networks often do overfitting on small size datasets, we adapted the VGG network based on the trauma dataset (Fig. 4). Our modification included adding a batch normalization [24] to the convolutional neural networks (CNN) to speed up the training operation and assist the regularization. We also used the dropout [25] and Gaussian noise to prevent overfitting and increase generalization. For the activation function, we used rectified linear units (ReLUs) and the last classification layer included the global average pooling followed by a softmax activation function to calculate the prediction probabilities.

B. Keyword and Fusion Networks

As shown in Fig 5, we designed an architecture that consists of a keyword network, an audio network, and a fusion network. We used a fully-connected network (FCN) layer with the ReLU activation function to generate the keyword feature representation. We empirically evaluated different sizes and number of layers to find the optimal network configuration (Table VIII). The results show that using a single FCN layer with size 128 achieved the best performance. Increased number of FCN layers (deeper) or the number of FCN layer units (wider) decreased the performance. The fusion network concatenated the audio network output features (a) and the keyword module outputted features (w) into one vector (y):

$$y = \gamma(concat(\phi(w),\ \psi(a))) \tag{1}$$

where ϕ, ψ and γ are the fully connected network layers (FCN), and y is the output of the fusion, which includes another FCN and ReLU activation function to generate the high-level feature

TABLE VII. THE ACCURACY AND AVERAGE F1-SCORE FOR DIFFERENT INPUT CHANNELS CONFIGURATIONS

Input channels	Accuracy	Average F1-score
CH1 only	22.2	0.106
CH2 only	22.9	0.121
United CH1 with CH2	22.6	0.115
(CH1 + CH2)/2	30.2	0.217
Combined CH1 & CH2	**30.8**	**0.231**

Input 60 × 40 × 20
5 × 5 CNN(128) + BN + ReLU 3 × 3 CNN(128) + BN + ReLU 2 × 2 Max-Pooling Gaussian-Noise(1.0)
3 × 3 CNN(256) + BN + ReLU 3 × 3 CNN(256) + BN + ReLU 2 × 2 Max-Pooling Gaussian-Noise(0.75)
3 × 3 CNN(512) + BN + ReLU Dropout(0.3) 3 × 3 CNN(512) + BN + ReLU Dropout(0.3)
3 × 3 CNN(512) + BN + ReLU Dropout(0.3) 3 × 3 CNN(512) + BN + ReLU Dropout(0.3) 2 × 2 Max-Pooling Gaussian-Noise(0.75)
3 × 3 CNN(1024) + BN + ReLU Dropout(0.5) 1 × 1 CNN(1024) + BN + ReLU Dropout(0.5)
1 × 1 CNN(6) + BN + ReLU Gaussian-Noise(1.0) Global-Average-Pooling
6-way Softmax

Fig. 4. Our audio network architecture. BN: Batch Normalization, ReLU: Rectified Linear Unit.

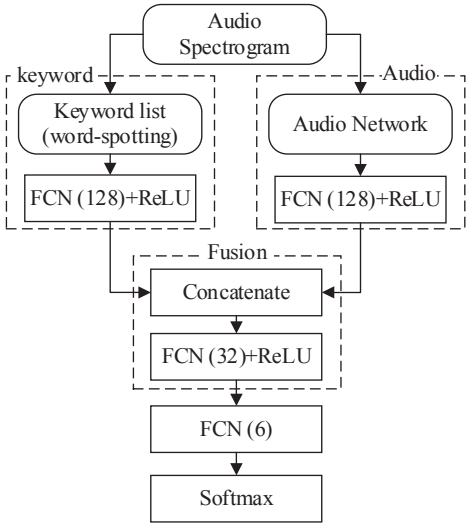

Fig. 5. Final model architecture after adding the keyword features. FCN: Fully Connected Network, ReLU: Rectified Linear Unit

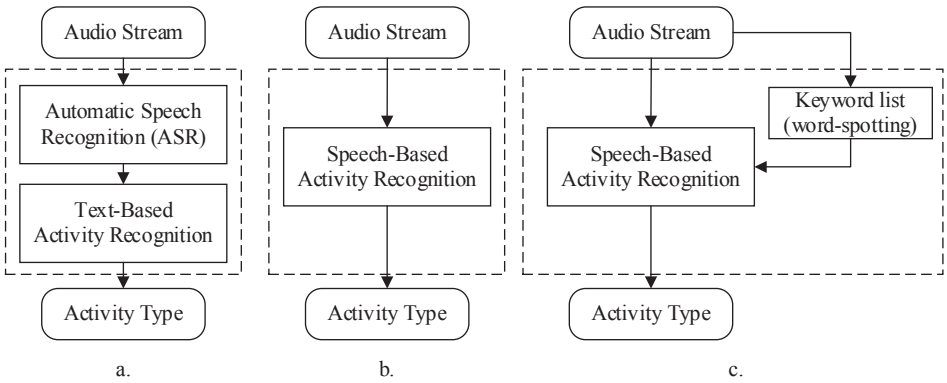

Fig. 6. Speech-based activity recognition proposed architectures. a. An architecture that uses the predicted transcripts from automatic speech recognition. b. An architecture that predicts the activity type directly from the audio. c. Same as in (b) with an additional one keyword input obtained from word-spotting.

representation for the final softmax layer classification (Fig. 5). We compared different fusion methods such as attention, but it did not perform well due to audio and keyword missalignment issues. This issue will be addressed in our future work.

VI. Experiments Setup and Results

We trained and evaluated the proposed model on the trauma resuscitation dataset. We used the utterances from the five most frequent activities and the total number of utterances was 3557. The dataset was randomly shuffled and split into 80% and 20% as a training set and a testing set, respectively. Each sample was considered independently, which contains an utterance-level audio stream, the related one keyword, and the correspondence activity label assigned by the trauma experts. Due to the small data size, we applied the fivefold cross-validation. We trained all networks together as end-to-end using early fusion approach. We use Adam [26] optimization with 0.001 as the learning rate and categorical crossentropy loss function. Each experiment took about two hours. We implemented all the experiments using Keras API of the TensorFlow library [27] with two Nvidia GTX 1080 GPUs.

Fig. 6 shows the diagrams of possible architectures for speech-based activity recognition. The first design (Fig. 6 (a)) integrates an automatic speech recognition (ASR) module with a text-based activity recognition (TAR) module. The overall performance of this design highly depends on both modules. Although the previous TAR model [6] achieved 69.1 accuracy and 0.67 average F1-score on a resuscitation dataset, their model used the human transcripts and assumed the ASR system can achieve human parity. Unfortunately, our ASR experimental results showed a high word error rate (WER) on the resuscitation dataset using two different architectures: attention-based seq2seq [28] and N-gram [29], which achieved WER of 100.3 and 75.8, respectively. We believe that the poor audio quality caused by the distant-talking, ambient noise, fast speaking rate, and concurrent speakers reduced the overall activity recognition performance, which made this model infeasible. The second architecture predicts the activity type directly from the audio (Fig. 6 (b)). Our evaluation result showed that the model achieved 30.8% in accuracy and 0.231 in average F1-score. Compared with the above two architectures, the proposed model (shown in Fig. 6 (c)) achieved 45.4 in accuracy and 0.415 in F1-score, which outperformed the previous approaches that used the audio directly or required the full utterance transcript. This comparison result demonstrates the strength of using the keyword as an additional feature to the speech-based activity recognition architecture.

We further compared our audio network with other state-of-the-art classification architectures such as VGG16-19 [7], DenseNet [22], ResNet [21] and NASNetMobile [30] (Table IX). The result showed that our audio network outperformed others in terms of accuracy and the average F1-score by a range of 1.3% – 8.9% and 0.02 – 0.129, respectively. This is because the general deep architectures usually suffer from overfitting when applied to the audio processing [18].

We made a quantitative analysis by comparing the performance of the three models using different inputs: audio-only, keyword only, and both audio and keyword (Table X). The result showed that using both audio and keyword features outperformed using audio-only or keyword only, which confirmed our hypothesis that keywords can boost the performance of the audio-only model, but not to replace it.

Table XI shows the F1-score for each activity by different modalities. The audio network had better performance on

TABLE VIII. Results Comparsion Between Different Keyworkd and Fusion Modules Layer Structure

Audio + Keyword	Accuracy %	Average F1-Score
(1-layer, 64)	44.9	0.412
Deeper (2-layers, 64)	44.9	0.409
Deeper (2-layers, 128)	44.8	0.409
Wider (1layer, 256)	44.7	0.409
(1-layer, 128)	**45.4**	**0.415**

TABLE IX. Results Comparsion Between Our Network and Other Classification Models

Classification Models	Accuracy %	Average F1-Score
NASNetMobile [24]	21.9	0.102
VGG19 [14]	27.7	0.182
DensNet [17]	28.2	0.190
ResNet [16]	28.0	0.196
VGG16 [14]	29.5	0.211
Our Network	**30.8**	**0.231**

TABLE X. RESULTS COMPARSION BETWEEN KEYWORD AND AUDIO MODELS

Modality Type	Accuracy %	Average F1-score
Audio only	30.8	0.231
Keyword only	38.3	0.344
Audio + Keyword	**45.4**	**0.415**

Extremity and *Back* activity than *GCS*, *Face* and *Circulation Control*. Different factors can cause these variations: imbalanced dataset and the noise level. As seen in Table II, the number of utterances that include each activity decreased by 100 from *Extremity* to *Circulation Control* revealing an unequal distribution between the activities. This imbalance caused the neural network classifiers to get biased towards the high-frequency activities more than low-frequency activities. As for the noise level, Fig. 1 shows that *GCS* and *Face* had relatively higher noise than other activities, which may impact the prediction performance. The third column of Table XI shows the F1-scores of each activity for the final model that fuses both the audio stream and keywords. The scores were boosted for almost all activities.

VII. CONCLUSIONS AND FUTURE WORK

We introduced a novel model for speech-based activity recognition and empirically evaluated it on a trauma resuscitation dataset. In our design, we extend the input features of the audio stream by integrating keywords—single-words from the most frequent words list associated with each activity. The new structure showed a substantial increase in the accuracy and the average F1-score 3.6% and 0.184, respectively, compared to the audio network alone. Due to the high word error rate of the ASR output caused by the fast speaking rate, concurrent speakers, and high ambient noise, our approach that relies single keywords instead of the entire ASR generated utterances is more efficient. We also analyzed the trauma resuscitation audio constraints, such as activity recurrence, noise level and most frequent words. In the evaluation results, we found that the imbalance of the activity frequencies in the trauma resuscitation, as well as the noise, reduced the accuracy of the audio network. Also, we explored audio stream preprocessing factors, such as different ways of combining the audio channels and features types. We found that the static MFSC features and the stereo channel configuration had the best performance. We introduce a new audio network based on the VGG model and provided an evaluation comparison with various classification architectures. Our model with relatively few layers, outperformed other classifiers .

Introducing the keyword features is promising, but further experiments on integrating the word-spotting models with the current architecture are needed for a more accurate evaluation. Also, we will evaluate more architectures for the fusion and keyword modules.

TABLE XI. THE F1-SCORE FOR EACH ACTIVITY FOR DIFFERENT MODALITIES

Activity	F1-score		
	Audio	Keyword	Audio + Keyword
Extremity	**0.366**	**0.532**	**0.524**
Back	**0.448**	**0.375**	**0.582**
GCS Calculation	0.124	0.314	0.313
Face	0.054	**0.389**	**0.385**
Circulation Control	0.045	0.242	0.255
OTHER	**0.351**	0.212	**0.429**

ACKNOWLEDGMENT

This research has been supported by the National Library of Medicine of the National Institutes of Health under grant number 2R01LM011834-05 and by the National Science Foundation under grant number IIS-1763827.

REFERENCES

[1] I. Chakraborty, A. Elgammal, and R. S. Burd, "Video based activity recognition in trauma resuscitation," IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–8, 2013.

[2] X. Li, Y. Zhang, I. Marsic, A. Sarcevic, and R. S. Burd, "Deep Learning for RFID-Based Activity Recognition," International Conference on Embedded Networked Sensor Systems, vol. 2016. pp. 164–175, 2016.

[3] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. J. M. Havinga, "Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey," ARCS Workshops. pp. 167–176, 2010.

[4] S. Jagannath, A. Sarcevic, N. Kamireddi, and I. Marsic, "Assessing the Feasibility of Speech-Based Activity Recognition in Dynamic Medical Settings," Human Factors in Computing Systems. 2019.

[5] X. Li et al., "Concurrent Activity Recognition with Multimodal CNN-LSTM Structure," CoRR, vol. abs/1702.0, 2017.

[6] Y. Gu et al., "Multimodal Attention Network for Trauma Activity Recognition from Spoken Language and Environmental Sound," IEEE International Conference on Healthcare Informatics. pp. 1–6, 2019.

[7] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations. 2015.

[8] S. Parlak and I. Marsic, "Detecting Object Motion Using Passive RFID: A Trauma Resuscitation Case Study," ieee Trans. Instrum. Meas., vol. 62, no. 9, pp. 2430–2437, 2013.

[9] S. Parlak, S. Ayyer, Y. Y. Liu, and I. Marsic, "Design and Evaluation of RFID Deployments in a Trauma Resuscitation Bay," ieee J. Biomed. Heal. informatics, vol. 18, no. 3, pp. 1091–1097, 2014.

[10] X. Li et al., "Activity recognition for medical teamwork based on passive RFID," International Conference on RFID. pp. 1–9, 2016.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," Computer Vision and Pattern Recognition. pp. 1725–1732, 2014.

[12] X. Li et al., "Region-based Activity Recognition Using Conditional GAN," ACM Multimedia, vol. 2017. pp. 1059–1067, 2017.

[13] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems 30, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

[14] B. Boashash, Time-frequency signal analysis and processing: a comprehensive reference. Academic Press, 2015.

[15] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCC extraction method in speech recognition," International Symposium on Circuits and Systems. 2006.

[16] M. F. McKinney and J. Breebaart, "Features for audio and music classification," International Symposium/Conference on Music Information Retrieval. 2003.

[17] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," ieee Trans. audio speech Lang. Process., vol. 22, no. 10, pp. 1533–1545, 2014.

[18] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, "The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification," European Signal Processing Conference. pp. 1–5, 2019.

[19] K. Kumar, C. Kim, and R. M. Stern, "Delta-spectral cepstral coefficients for robust speech recognition," International Conference on Acoustics, Speech, and Signal Processing. pp. 4784–4787, 2011.

[20] S. Hershey et al., "CNN architectures for large-scale audio classification," International Conference on Acoustics, Speech, and Signal Processing. pp. 131–135, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Computer Vision and Pattern Recognition. pp. 770–778, 2016.

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Computer Vision and Pattern Recognition. pp. 2261–2269, 2017.

[23] H. Eghbal-zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," European Signal Processing Conference. pp. 2749–2753, 2017.

[24] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," Feb. 2015.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[26] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations. 2015.

[27] M. Abadi et al., "Tensorflow: a system for large-scale machine learning.," in OSDI, 2016, vol. 16, pp. 265–283.

[28] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4960–4964.

[29] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," Conference of the International Speech Communication Association. 2002.

[30] B. Zoph, V. Vasudevan, J. Shlens, and Q. V Le, "Learning Transferable Architectures for Scalable Image Recognition," Computer Vision and Pattern Recognition. pp. 8697–8710, 2018.